Master 2 Traitement Automatique des Langues
parcours Recherche & Développement

Fouille de résumés d'articles biomédicaux

Annotation d'entités et extraction de relations

Juliette POTIER
potierjuliette@yahoo.fr

Directrice :

Isabelle Tellier
Université Paris 3
LATTICE
isabelle.tellier@sorbonne-nouvelle.fr

Rapporteur externe :

Mathieu Valette
INALCO – ER-TIM
mvalette@inalco.fr

Encadrants :

Thierry Hamon
Université Paris 13 – LIMSI
thierry.hamon@limsi.fr

Natalia Grabar
Université Lille 3 – STL
natalia.grabar@univ-lille3.fr

CNRS – UPR 3251 Laboratoire Informatique pour la Mécanique et les Sciences de l'Ingénieur

6 septembre 2017

Table des matières

Table des figures

Liste des tableaux

Remerciements

Je tiens à remercier toutes les personnes qui ont aidé à la rédaction de ce mémoire. Mes remerciements s'adressent en particulier à Thierry Hamon et Natalia Grabar qui ont encadré ce travail pendant mon stage au LIMSI ainsi qu'à Isabelle Tellier, Serge Fleury et Cyril Grouin. Je remercie enfin l'ensemble de mes collègues de bureau pour leur aide et bonne humeur, Lionel, Lauriane, Judith, Mahmut ainsi que l'ensemble des stagiaires, doctorants et post-doctorants du LIMSI.

Résumé

Le présent mémoire décrit deux étapes de fouille d'articles scientifiques du domaine biomédical sur la thématique des relations aliments - médicaments. Nous avons dans un premier temps procédé à l'évaluation de ressources pour l'annotation des entités en jeux dans les relations aliments - médicaments (ci-nommées entités biomédicales) et dans un second temps entraîné des modèles de classification des relations à partir des contextes et des arguments de ces dernières. Nous avons pour cela utilisé en point de départ le corpus POMELO annoté manuellement avec Brat en relations entre entités biomédicales. En utilisant des ressources existantes, nous avons ainsi projeté les termes de ces ressources sur le corpus POMELO avant de soumettre les annotations à une évaluation par rapport aux annotations de référence avec BratEval. Nous avons par ailleurs extrait les relations annotées du corpus POMELO au moyen de scripts Perl, puis entraîné les modèles et comparé les performances de chacun avec le logiciel Weka. L'annotation des entités obtient globalement de meilleurs résultats lorsque l'on complète les ressources mais les scores en macro-évaluation restent inférieurs à 0.3 pour la précision et à 0.2 pour le rappel et la F-mesure. La classification des relations tourne autour de 0.7 en termes de macro-précision, macro-rappel et F-mesure et autour de 0.9 pour certains types de relations.

Mots clés : *Projection de ressources, Patrons d'extraction, Classification de relation, Extraction de relation, Apprentissage automatique*

Chapitre 1

Introduction

Le travail présenté dans ce mémoire s'inscrit dans le projet ANR MIAM coordonné par le LIMSI (CNRS) [1]. L'un des objectifs de ce projet est d'extraire les relations aliments-médicaments décrites dans la littérature biomédicale afin d'enrichir les bases de données existantes. L'extraction de relation est, de manière plus générale, un domaine découlant de l'extraction d'information.

L'enjeu principal de l'extraction d'information est de rechercher des informations contenues dans des données trop volumineuses pour être parcourues manuellement. Cela est particulièrement observable dans le domaine biomédical où la quantité d'articles scientifiques ne cesse d'augmenter [Zhou et al., 2014, Charnois et al., 2009]. Disposer de connaissances à jour par rapport à la recherche scientifique peut s'avérer impératif pour les professionnels de santé. Beaucoup de ces informations se trouvent par ailleurs dans des bases de données [2]. Parmi les sujets d'études de ce domaine, on note entre autre l'étude des relations entre l'alimentation et les médicaments [Jovanovik et al., 2015].

En effet, les interactions aliments-médicaments peuvent aboutir à des effets dangereux pour la santé des patients [Grabar et al., 2015]. Dans certain cas, l'absorption d'un aliment peut interagir avec le métabolisme d'un médicament en diminuant ses effets, en les augmentant ou en provocant un effet secondaire indésirable. Par exemple, manger du pamplemousse en parallèle d'une prise de

1. `https://miam.limsi.fr/`
2. `http://lod-cloud.net/`

certains médicaments peut altérer le métabolisme de ces derniers et provoquer un surdosage.

Le champ de recherche est donc pluridisciplinaire et fait intervenir des méthodes relevant à la fois du domaine du Traitement Automatique des Langues et de l'apprentissage automatique. En partant de l'hypothèse que les relations apparaissent dans des contextes différents en fonction de la relation qu'ils décrivent, on peut donc estimer qu'apprendre à détecter ces contextes peut permettre de déduire la nature des relations.

Le mémoire se focalisera dans un premier temps sur l'annotation des entités en jeu dans l'expression d'une interaction aliment-médicament. Elles regroupent les maladies, les médicaments, les aliments, les effets secondaires et toutes les informations en rapport avec l'administration d'un médicament. Ces entités seront dans la suite du document désignées sous le terme d'*entités biomédicales*. Le but recherché sera d'évaluer les ressources décrites au chapitre 3 afin de déterminer si celles-ci suffisent à une annotation en entités biomédicales. Dans un second temps, nous aborderons la question de la classification des relations du corpus de citations d'articles scientifiques.

Nous nous pencherons principalement dans l'état de l'art sur les méthodes d'extraction de relations et d'apprentissage automatique de patrons d'extraction, nous décrirons ensuite le corpus POMELO et les ressources constituant les données sur lesquelles nous allons travailler. Nous détaillerons les méthodes choisies pour l'annotation des entités et pour la classification des relations et enfin nous discuterons les résultats des différentes configurations par rapport aux annotations du corpus POMELO et par rapport à l'état de l'art.

Chapitre 2

État de l'art

1 Introduction

Nous allons focaliser l'état de l'art sur l'extraction de relations. En effet, l'annotation des entités biomédicales se fera avec la projection des ressources détaillées au chapitre 3 section 3. Le but est d'évaluer la pertinence de ces dernières pour l'annotation des relations entre aliments et médicaments. Nous ne nous pencherons donc pas sur l'état de l'art de l'annotation d'entités dans des corpus d'articles biomédicaux.

2 Extraction de relation

L'extraction de relations est une branche de l'extraction d'information. On peut la définir, selon [Bach and Badaskar, 2007], comme l'extraction de « tuples $t = (e_1, ..., e_n)$ où e_i sont les entités d'une relation prédéfinie ». La majeur partie des travaux actuels se focalisent sur l'extraction de relations binaires [Bach and Badaskar, 2007].

L'ensemble des travaux actuels en terme d'extraction de relations dans le domaine biomédical fait état de la nécessité à améliorer les méthodes existantes pour faciliter l'accès aux informations. Tous les types d'interactions sont donc susceptibles d'être étudiés en particulier ceux faisant intervenir l'absorption d'un élément assimilable par l'organisme (médicaments, compléments alimentaires, aliments, etc) comme les interactions entre

médicaments [Abacha et al., 2015, Tatonetti et al., 2012] et [Liu et al., 2016], entre médicaments et maladies, etc. D'autres études du domaine de la bio-informatique se sont entre autres penchées sur l'extraction des relations entre protéines [Huang et al., 2004] et sur l'extraction d'événements dans la littérature biomédicale [Nguyen and Grishman, 2015].

L'extraction de relations est actuellement constituée de deux approches différentes : le TAL symbolique (*rule-based methods*) d'une part qui se base sur l'utilisation de grammaires et les méthodes statistiques (*machine learning based methods*) d'autre part qui se basent sur de l'apprentissage automatique et des méthodes de classification supervisée [Cohen and Hunter, 2008]. Il existe cependant d'autres approches pour l'extraction de relations moins détaillées dans la littérature à l'exemple de l'exploitation des phénomènes de co-occurence [Wei et al., 2016].

2.1 TAL symbolique

[Charnois et al., 2009] décrivent les méthodes symboliques en citant (Zweigenbaum *et al.*, 2007) comme s'appuyant sur « [des] lexiques, [des] règles d'extraction, [des] analyse[s] syntaxique[s] voire sémantique[s] de la phrase ». Elles ont l'avantage d'être facilement interprétables par les chercheurs car plus compréhensibles pour les non-experts mais donnent cependant des résultats limités en terme de rappel même si la précision des méthodes basée sur les règles d'extraction est excellente. [Charnois et al., 2009] citent notamment les travaux de (Fukuda *et al.*,1998) et (Hakenberg *et al.*, 2008) pour l'utilisation d'expressions régulières, (Tsuruoka et ichi Tsujii, 2003) pour l'utilisation de dictionnaires, les travaux de (Tsai *et al.*, 2006) pour les méthodes de d'analyses syntaxiques, et les travaux de (Ono *et al.*, 2001 ; Ng et Wong, 1999) pour des méthodes mixtes. [Huang et al., 2004], qui exploitent entre autre un lexique de verbes déclencheurs de relation entre deux entités, citent notamment (Leroy et Chen, 2002) pour leur méthode d'analyse focalisée sur les prépositions.

2.2 Apprentissage automatique

Les méthodes statistiques ou probabilistes sont les méthodes d'extraction de relations qui obtiennent les meilleurs résultats [Charnois et al., 2009]. Elles utilisent diverses méthodes telles que les machines à vecteurs de support (*Support Vector Machines*) chez [Aramaki et al., 2010] qui utilise en traits les lexiques des entités ciblées ainsi que le contexte et la distance entre celles-ci. Les modèles de Markov cachés (*Hidden Markov Models*), les champs aléatoires conditionnels (*Conditional Random Fields*), les arbres de décision (*Decision Tree*) [Fundel et al., 2007] et les réseaux neuronaux convolutifs (*Convolutional Neural Networks*) [Bach and Badaskar, 2007, Liu et al., 2016, Jiang et al., 2016, Zhou et al., 2014] sont d'autres moyens d'apprendre automatiquement des relations. [T. et al., 2009] utilise ces derniers avec des traits sémantiques, syntaxiques et de position. Les méthodes à noyau (*kernel methods*) [Zelenko et al., 2003] et (Krallinger et al. 2008) sont un peu à l'écart des méthodes précédentes en se distinguant des autres par le fait qu'elles utilisent moins d'ensembles de traits en entrée [Zhou et al., 2014].

Même si elles tendent à obtenir de meilleurs résultats, les méthodes d'apprentissage automatique restent néanmoins interprétables uniquement par des spécialistes. Elles ont par ailleurs l'inconvénient de ne pas être exploitables si l'on ne possède pas d'exemples et d'être difficilement extensibles à l'extraction de relations non-binaires.

2.3 Méthodes hybrides

Les méthodes du TAL symbolique et de l'apprentissage automatique ne sont pas antagonistes. Elles tendent à se combiner dans les recherches actuelles comme celles décrites dans [Wei et al., 2016]. *BioCreative V chemical-disease relation task*, organisé par [Wei et al., 2016], est un challenge qui vise la réalisation de deux tâches principales : la reconnaissance des noms de maladies d'une part, et l'extraction des relations entre produits chimiques et maladies d'autre part. Les organisateurs ont rassemblé un corpus annoté selon les entités (produits chimiques et maladies) et relations recherchées à partir d'articles issus de PubMeb. 34 équipes de recherches ont participé aux tâches

sus-mentionnées et ont majoritairement proposé des systèmes combinant TAL symbolique et apprentissage automatique. Les résultats combinés des différents systèmes proposés ont permis d'obtenir une F-mesure de 0.8889 pour les systèmes de reconnaissance des noms de maladies et de 0.6280 pour les systèmes d'extraction de relations.

3 Apprentissage de patrons d'extraction

L'apprentissage automatique de patrons d'extraction de relations est depuis plusieurs dizaines d'années un domaine actif de la recherche en Traitement Automatique des Langues. [Hearst, 1992] proposait déjà en 1992 une méthode d'apprentissage de patrons pour extraire des relations d'hyponymie. Elle repose sur une analyse lexico-syntaxique des textes à partir de laquelle on extrait de nouvelles paires d'entités reliées par une relation d'hyponymie. Le système PROMÉTHÉE [Morin, 1999] permet d'extraire des patrons lexico-syntaxiques de manière incrémentale à partir de tuples « relation, terme 1, terme 2 » dont les attributs sont connus. Le processus itératif pour extraire des patrons de relations est ensuite largement employé [Jean-Louis et al., 2011], les travaux de ce domaine concernant des textes de tout domaines qu'ils soient en dehors du domaine biomédical [Riloff, 1996] et [Cellier and Charnois, 2010] ou bien spécialisés dans celui-ci [Meng and Morioka, 2015, Huang et al., 2004, Aussenac-Gilles and Condamines, 2009].

4 Conclusion

Les approches traditionnelles en extraction de relation sont les méthodes du TAL symbolique qui utilisent des lexiques et des patrons d'extraction et de l'apprentissage automatique qui utilisent des ensembles de traits sur lesquels se base l'apprentissage des relations. Plus récemment, ces deux approches se combinent notamment en proposant des méthodes d'apprentissage automatique de patrons d'extraction.

Chapitre 3

Matériel

1 Introduction

Le matériel que nous allons utiliser a été constitué lors du projet POMELO [Grabar et al., 2015]. Il est constitué d'un corpus de textes annotés en relations et de ressources terminologiques rassemblées en vu de l'annotation automatique du corpus en entités biomédicales.

2 Le corpus POMELO

Le corpus a été constitué lors du projet du même nom [Grabar et al., 2015] et est composé de titres et de résumés d'articles scientifiques biomédicaux en anglais. Il comporte 639 fichiers textes et 639 fichiers d'annotations en entités biomédicales et en relations aliments-médicaments-maladies. Les titres et résumés du corpus ont été collectés depuis le portail PubMed [1] avec la requête « ("FOOD DRUG INTERACTIONS"[MH] OR "FOOD DRUG INTERACTIONS*") AND ("adverse effects*") » afin de rassembler des résumés mentionnant des interactions entre aliments et médicaments ainsi que les effets provoqués par ces dernières.

Le tableau 3.1 résume les caractéristiques générales du corpus dont la taille est d'environ 117 000 mots pour 13 694 entités et 2 486 relations annotées. Un exemple de fichier annoté est donnée en figure 3.1. L'annotation du corpus

1. `https://www.ncbi.nlm.nih.gov/pubmed/`

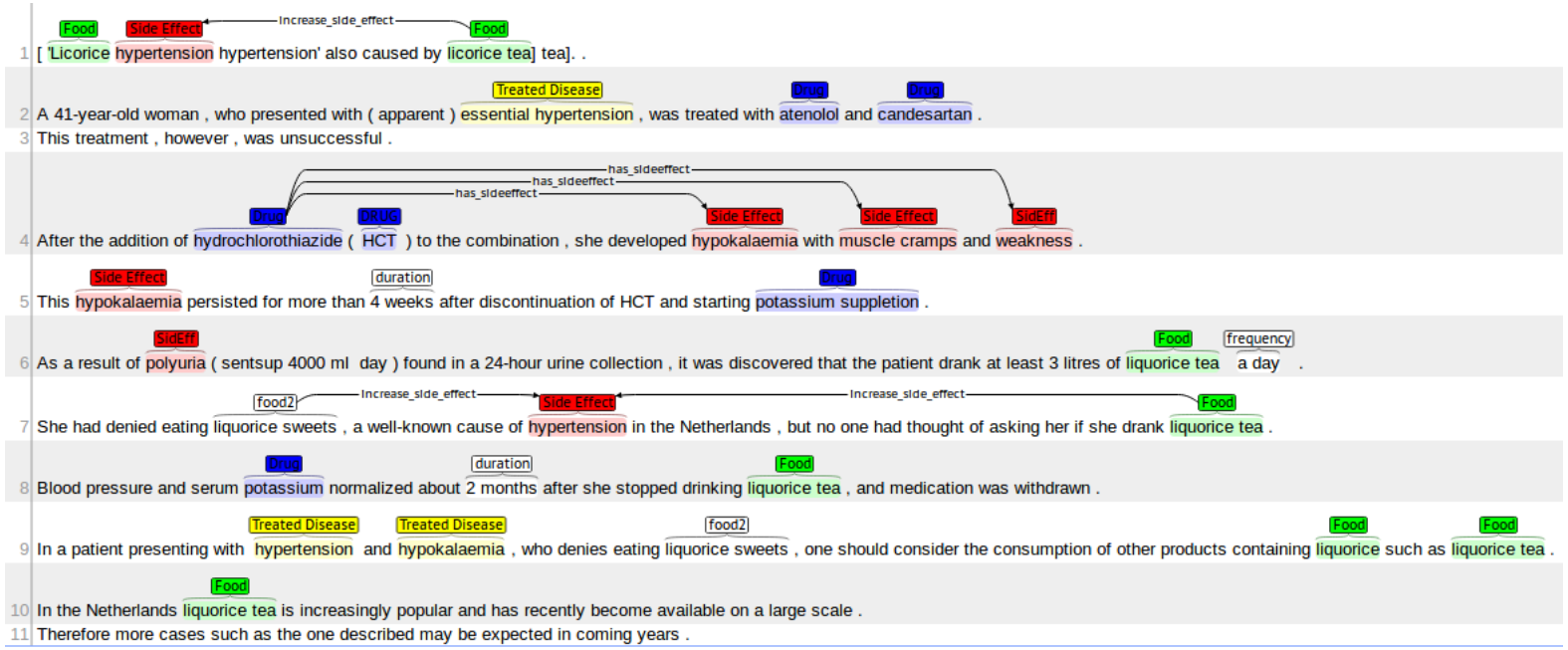
a été effectuée manuellement par Vincent Tabanou, interne en pharmacie, au moyen du logiciel Brat [2]. Les tableaux 3.2 et 3.4 présentent plus en détails les 17 types d'entités et 21 types de relations qui ont été définies pour l'annotation ainsi que les fréquences de ces étiquettes sémantiques.

TABLEAU 3.1 – Corpus POMELO

	Nb total	Moyenne par fichier	min. par fichier	max. par fichier
mots	116 503	182	1	699
entités	13 694	21	1	116
relations	2 486	4	0	33

2. http://brat.nlplab.org/

Figure 3.1 – Exemple d'un résumé annoté



2.1 Entités biomédicales

Dans le tableau 3.2, 17 étiquettes sémantiques regroupent des entités qui partagent des caractéristiques sémantiques communes. Les aliments de base tels que les fruits et les légumes qui peuvent être cuisinés en des plats plus complexes ou bien directement ingérés sont ainsi regroupés sous l'étiquette `food` (ex : 'grapefruit', 'water', 'skim milk') comme dans l'exemple donné en figure 3.2. Les plats cuisinés (ex : 'pizza', 'pepperoni', 'snack') sont désignés par l'étiquette `food2` et les termes plus génériques en rapport avec l'alimentation et la cuisine (ex : 'breakfast', 'nutrition') sont, quant à eux, regroupés sous les étiquettes `food3` et `food4`.

FIGURE 3.2 – Exemple d'entité annotée

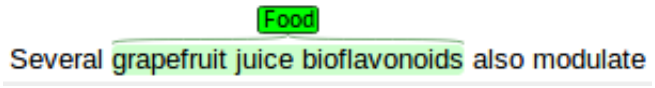


Tableau 3.2 – Corpus POMELO – Entités annotées

Entités	Étiquette	Nb d'occurences	Exemples
médicament et classe de médicaments	`drug`	4934	'chemotherapy'
aliment de base	`food`	2443	'water', 'milk'
effet indésirable	`sideeffect`	1985	'malnutrition', 'steatosis'
moment d'administration, termes «fed/fasted»	`mealTime`	1027	'postprandial'
dosage	`dosage`	767	'10 mg'
maladie	`treatedDisease`	645	'depression', 'malaria'
effet du médicament, pharmacodynamique	`drugEffect`	558	'bioavailability'
mode d'administration	`modeadm`	539	'oral'
fréquence de prise ou posologie	`frequency`	282	'once daily'
complément alimentaire	`foodSupplement`	186	'calcium'
durée du traitement	`duration`	86	'for 7 days'
nourriture composée	`food2`	77	'pepperoni'
terme en lien avec l'alimentation et la cuisine	`food3`	72	'breakfast'
nombre de sujets d'une étude	`numbers`	59	'12', 'Five patients'
abandonné en cours d'annotation	`drug2`	19	
autre	other	12	'metabolism'
identique à `food3`	food4	3	
TOTAL	17	13 694	

On remarque que les entités les plus présentes dans le corpus sont les médicaments, les aliments et les effets secondaires. Viennent ensuite les moments d'administration et les dosages.

2.2 Interactions aliments-médicaments

Les relations du corpus se focalisent sur les interactions entre les aliments, les médicaments et les maladies, plus particulièrement celles provocant des effets indésirables ou bien des effets sur le métabolisme médicamenteux. Le tableau 3.3 répertorie le nombre de relations annotées en fonction des événements observés par les chercheurs. Les figures 3.3, 3.4 et 3.5 présentent des exemples de relations annotées pour une absence d'interaction, une interaction ralentissant l'élimination du médicament par l'organisme et une interaction entre un médicament, une maladie et un effet secondaire.

TABLEAU 3.3 – Thématiques

Relation concernant	Nb d'occurences
effets secondaires	696
effets sur les médicaments	259
absorption	141
effets médicamenteux	30
indication de temps	26
élimination	18

FIGURE 3.3 – Relation `no_effect_on_drug`

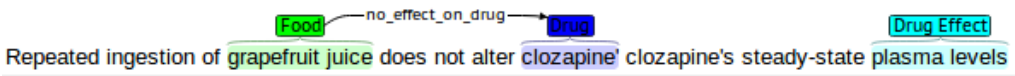


FIGURE 3.4 – Relation `slow_elimination`

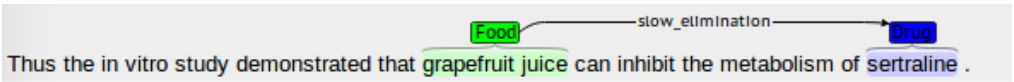


FIGURE 3.5 – Relations `treat` et `has_sideeffect`

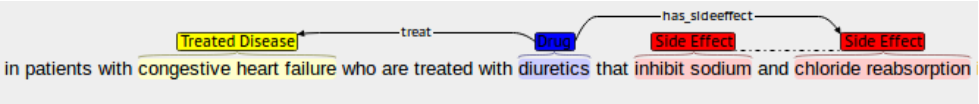


TABLEAU 3.4 – Corpus POMELO – Relations annotées

Relation	Étiquette	Occurrences
relation non précisée	`relation`	706
a un effet indésirable	`has_sideeffect`	434
traite, soigne	`treat`	350
exacerbe l'effet indésirable	`increase_side_effect`	239
a un effet	`has_effect`	233
aucun effet sur le médicament	`no_effect_on_drug`	145
effet négatif sur le médicament	`negative_effect_on_drug`	91
diminue l'absorption du médicament	`decrease_absorption`	64
augmente l'absorption	`increase_absorption`	52
doit être pris sans aliment	`without_food`	23
effet positif sur le médicament	`positive_effect_on_drug`	23
atténue l'effet indésirable	`reduce_side_effect`	23
ralentit l'absorption	`slow_absorption`	21
ralentit l'élimination	`slow_elimination`	18
diminue l'action médicamenteuse	`worsen_drug_effect`	16
augmente l'action médicamenteuse	`improve_drug_effect`	14
après le repas	`after`	11
pendant le repas	`during`	10
avant le repas	`before`	5
apparition d'un effet indésirable	`new_side_effect`	4
accélère l'absorption	`speed_up_absorption`	4
accélère l'élimination	`speed_elimination`	0
TOTAL	22	2 486

22 types de relations ont été définies pour annotation mais l'une d'entre elle n'a aucune occurrence dans le corpus. Une grande partie des relations ont été regroupées sous une appellation générique et une grande partie des relations spécifiées concernent les relations entre un médicament et une maladie ou bien des effets indésirables.

3 Ressources

En complément du corpus, nous avons utilisé les bases de données répertoriées pendant le projet POMELO [Grabar et al., 2015]. Il s'agit des ressources suivantes :

— USDA (National Nutrient Database for Standard Reference) [3] dédiée aux aliments et aux nutriments

— MESH [4] dédiée aux maladies et aux aliments

— ATC & DDD [5] pour les maladies et médicaments

— PharmGKB [6] dédiée aux maladies et médicaments

— RxNorm [7] dédiée aux médicaments

— Drugbank [8] dédiée aux médicaments

— MedDRA [9] dédiée aux maladies

— UMLS [10] Unified Medical Language System

— SIDER [11] dédiée aux effets secondaires

— Diseasome [12] dédiée aux maladies

3. https://ndb.nal.usda.gov/ndb/search/list
4. http://www.nlm.nih.gov/mesh/
5. http://www.whocc.no/ et http://www.genome.jp/kegg-bin/get_htext?query=08902&htext=br08902.keg
6. http://www.pharmgkb.org/
7. https://www.nlm.nih.gov/research/umls/rxnorm/
8. http://www.drugbank.ca
9. https://www.meddra.org/
10. https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html
11. http://sideeffects.embl.de
12. https://datahub.io/dataset/fu-berlin-diseasome

Dans le but de constituer des ressources terminologiques, les termes désignant des entités biomédicales ont été récupérés depuis les bases de données et mis en correspondance avec une étiquette sémantique utilisée lors de l'annotation manuelle du corpus (voir tableau 3.5). À partir de ces bases de données, il n'a pas été possible de créer une ressource terminologique pour les effets médicamenteux (`drugEffect`). Nous avons donc pu constituer des ressources de taille hétérogène pour les étiquettes suivantes :

- `drug` (349 304 entités)
    - ⋆ Drugbank (86 154 entités)
    - ⋆ RxNorm (245 203 entités)
    - ⋆ PharmGKB (11 718 entités)
    - ⋆ ATC & DDD (6 229 entités)
- `disease` (877 087 entités)
    - ⋆ MedDRA (75 172 entités)
    - ⋆ UMLS (798 274 entités)
    - ⋆ Diseasome (3 641 entités)
- `food` (12 125 entités)
    - ⋆ USDA (11 328 entités)
    - ⋆ Mesh (797 entités)
- `foodSupplement` (153 entités)
    - ⋆ USDA
- `sideeffect` (19 244 entités)
    - ⋆ PharmGKB (10 097 entités)
    - ⋆ SIDER (9147 entités)

TABLEAU 3.5 – Correspondances des étiquettes entre les ressources et le corpus

Étiquettes du corpus	Ressource(s) exploitée(s)	Appellations dans les ressources
`food`	USDA	'food'
`food2`	USDA	'food2'
`food3`	USDA	'food3'
food4	non annoté	non annoté
`foodSupplement`	USDA	'nutrient'
`drug`	Drugbank	'drug'
	PharmGKB	-
	ATC & DDD	-
	RxNorm	'Pharmacologic Substance'
		'(Organic|Inorganic) Chemical'
		'Antibiotic'
		'Amino Acid Peptide, or Protein'
		'Indicator, Reagent, or Diagnostic Aid'
		'Hormone'
		'Enzyme'
		'Immunologic Factor'
drug2	non annoté	non annoté
`sideeffect`	SIDER	-
	MedDRA	-
	PharMAgkb	-
drugEffect	non annoté	non annoté
`treatedDisease`	MedDRA	-
	UMLS	'DISO'
	Diseasome	'diseasome'
mealTime	-	-
modeadm	-	-
frequency	-	-
dosage	-	-
duration	-	-
numbers	non annoté	non annoté
other	non annoté	non annoté

L'utilisation des bases de données pour créer des ressources terminologiques pose le problème de l'attribution d'une étiquette plutôt qu'une autre : 'malnutrition' peut être considéré, selon le contexte, comme une maladie à soigner ou bien comme un effet secondaire. Les ressources issues de la base de données MedDRA ont donc été attribuées à la fois à l'étiquette `treatedDisease`

et à l'étiquette `sideeffect`.

4 Conclusion

Nous disposons donc d'un corpus de résumés d'articles biomédicaux annotés en entités et en relations aliments-médicaments-maladies ainsi que de plusieurs ressources terminologiques qui correspondent à plusieurs étiquettes d'annotation du corpus. Nous allons donc exploiter ces dernières afin d'évaluer leur pertinence pour l'annotation automatique d'un corpus de textes biomédicaux en utilisant les annotations manuelles comme référence. Nous allons par la suite exploiter les relations annotées pour déterminer les traits les plus pertinents pour les classifier en comparant plusieurs configurations.

Chapitre 4

Méthodes

1 Introduction

Après avoir décrit le corpus et les ressources que nous allons utiliser, nous allons maintenant détailler les méthodes d'annotation des entités et d'extraction des relations.

2 Prétraitements

Les fichiers texte du corpus ont été pré-traités et annotés avec la plateforme OGMIOS [Hamon and Nazarenko, 2008]. Les titres et résumés ont été segmentés en mots et en phrase et les termes ont ensuite été étiquetés morpho-syntaxiquement au moyen de TreeTagger [Schmid, 1994]. Pour la classification des relations, l'étiquetage morpho-syntaxique du texte a été effectué après l'extraction du contexte d'apparition des relations.

3 Annotation des entités biomédicales

L'annotation du corpus a été effectuée au moyen de projection de ressources terminologiques et de patrons d'extraction. L'ensemble de la chaîne de traitement, des prétraitements à l'évaluation, est visible en figure 4.1. Les entités biomédicales relatives à l'administration d'un produit ont été annotées avec une méthode de projection de patrons développée par

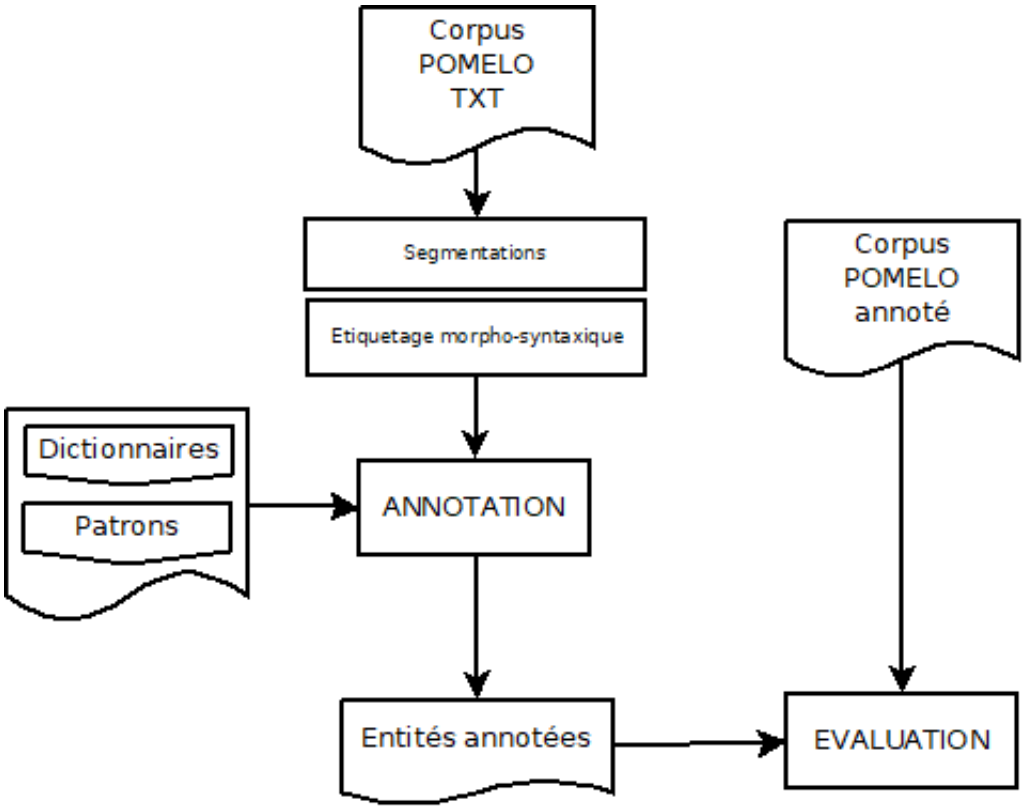
[Hamon and Grabar, 2010]. En effet, elles ne disposent pas ressources terminologiques. Les expressions régulières ont cependant été modifiées pour permettre une reconnaissance des moments d'administration (`mealTime`). Les autres entités ont été annotées au moyen des ressources terminologiques avec TermTagger [1] à l'exception des étiquettes `numbers` et `other` qui n'ont pas été prises en compte dans l'annotation en raison du faible nombre d'occurrence dans le corpus. Le tableau 4.1 répertorie les méthodes d'annotation des entités biomédicales en fonction de l'étiquette sémantique.

TABLEAU 4.1 – Annotation des entités

Étiquettes	Mode d'annotation
food	ressources terminologiques
food2	ressources terminologiques
food3	ressources terminologiques
food4	non annotée
foodSupplement	non annotée puis ressources
drug	ressources terminologiques
drug2	non annotée
sideeffect	ressources terminologiques
drugEffect	non annotée
treatedDisease	ressources terminologiques
mealTime	RegEx
modeadm	RegEx
frequency	RegEx
dosage	RegEx
duration	RegEx
numbers	non annotée
other	non annotée

1. `http://search.cpan.org/~thhamon/Alvis-TermTagger/`

FIGURE 4.1 – Processus d'annotation



4 Extraction et classification des relations

4.1 Choix de l'algorithme

Après l'annotation des textes à l'aide des ressources, nous nous sommes intéressés à l'extraction des relations sémantiques entre les entités biomédicales. Nous avons adopté une approche avec apprentissage automatique au moyen de Weka [Frank et al., 2016], un logiciel permettant d'appliquer un certain nombre d'algorithmes d'apprentissage sur un ensemble de données développé par l'université de Waikato [2]. Pour visualiser les critères de choix effectués par les modèles, nous avons employé l'algorithme C4.5 (J.R. Quinlan, 1996) implémenté sous le nom de J48. L'ensemble du processus de classification est présenté par la figure 4.2.

2. `http://www.cs.waikato.ac.nz/ml/weka/`

4.2 Constitution des datasets

Le contexte d'une relation est défini comme étant la fenêtre comprenant (x tokens, argument 1, tokens intermédiaires, argument 2, x tokens) dont un exemple est donné dans le tableau 5.4. Les contextes dans lesquels les relations aliments-médicaments apparaissent ont donc été extraits à partir des annotations Brat des arguments 1 et 2 et des fichiers texte au moyen de scripts Perl [3]. Les tuples « relation, premier argument, étiquette sémantique, deuxième argument, étiquette sémantique, contexte » servent de base à la création des fichiers ARFF. La classe d'apprentissage prend pour valeur une des 22 étiquettes de relation du corpus tandis que les attributs correspondant aux arguments de la relation prennent pour valeur une des 17 étiquettes sémantiques ayant servies à l'annotation des entités biomédicales.

3. `https://www.perl.org/`

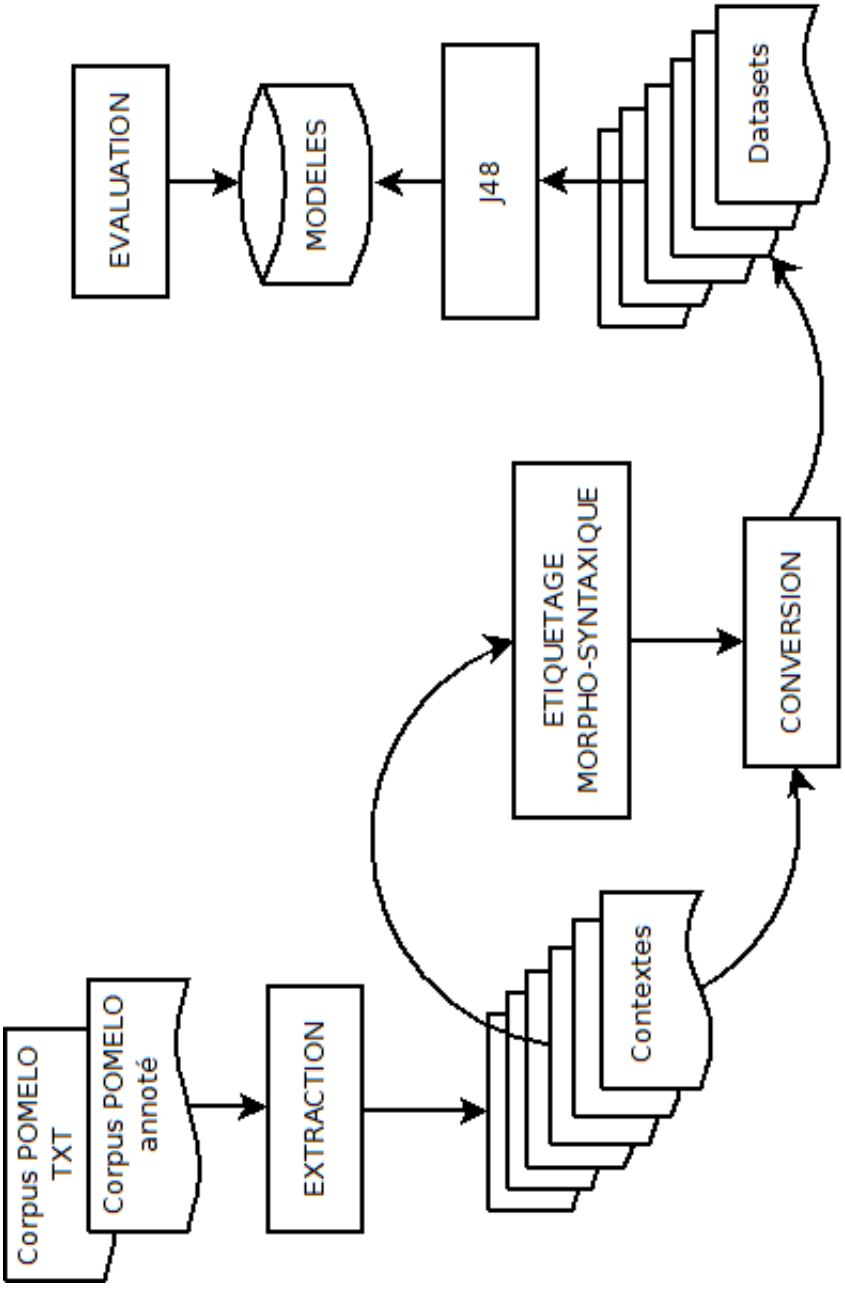


Figure 4.2 – Classification des relations

Les fichiers ARRF correspondant aux différents *datasets* contiennent tous 2486 instances qui correspondent aux relations annotées dans le corpus POMELO. Ils sont représentés par des données éparses sous la forme :

attribut1 valeur1, attribut2 valeur2, ... , étiquette1, étiquette2, relation

Les deux avant-derniers attributs correspondent aux étiquettes sémantiques des arguments de la relation ; le dernier attribut correspond à la classe d'apprentissage. Cela donne par exemple :

{159 1, 832 2, 3796 food, 3797 drug, 3798 relation}
{34 3, 35 4, 159 2, 2126 1, 3796 drug, 3797 food, 3798 relation}
{2 5, 21 3, 62 4, 1027 6, 1157 1, 2386 3, 3041 2, 3796 treatedDisease, 3797 drug, 3798 treat}

5 Mesures d'évaluation

Les réponses des systèmes ont été évaluées en terme de précision, rappel et F-mesure. Les deux premières mesures d'évaluation peuvent être calculer en micro ou en macro évaluation [Sebastiani, 2001]. Dans le premier cas, équations (4.1) et (4.2), le rappel et la précision sont calculés sans faire de distinction entre les résultats de chaque catégorie. Dans le second cas, équations (4.3) et (4.4), la précision et le rappel sont d'abord évalués localement pour chaque catégorie. On obtient la précision et le rappel globaux après avoir calculé la moyenne des précisions et rappels obtenus. La F-mesure, équation (4.5), est, quant à elle, la moyenne harmonique de la précision et du rappel.

$$P_\mu = \frac{TP}{TP + FP} \tag{4.1}$$

$$R_{\mu} = \frac{TP}{TP + FN} \tag{4.2}$$

$$P_M = \frac{\sum_{i=1}^{|C|} P_i}{|C|} \tag{4.3}$$

$$R_M = \frac{\sum_{i=1}^{|C|} R_i}{|C|} \tag{4.4}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{4.5}$$

Les réponses du système pour l'annotation des entités ont été évaluées au moyen de BratEval [4] en utilisant les annotations de POMELO comme gold standard. Les résultats étant donnés pour chaque étiquette, nous avons calculé la macro-évaluation de l'ensemble des annotations. Les évaluations de la classification ont été fournies par Weka en macro-évaluation avec les détails de chaque classe.

6 Conclusion

Nous avons effectué deux tâches avec deux méthodes différentes. Nous avons, dans un premier temps, annoté les entités biomédicales d'un corpus de résumés d'articles scientifiques en projetant des ressources terminologiques et des patrons d'extraction. Dans un second temps, nous avons extrait les relations annotées manuellement dans POMELO pour convertir les contextes de ces dernières en ensembles ARFF servant de bases pour l'apprentissage avec J48 de modèles de classification.

4. National ICT Australia Limited (NICTA) Biomedical Informatics, BRAT-Eval v1.0 (2013) – `https://bitbucket.org/nicta_biomed/brateval/`

Chapitre 5

Expérimentations et résultats

1 Introduction

Nous allons effectuer plusieurs expérimentations lors de l'annotation des entités en faisant varier les ressources. Nous allons aussi faire varier les descripteurs des ensembles ARRF pour entraîner des modèles de classification de relations en fonction des traits suivants : formes fléchies, lemmes ou étiquettes morpho-syntaxiques présentes dans le contexte en terme de fréquence ou de position en fonction d'une fenêtre plus ou moins grande.

2 Annotation des entités

2.1 Expérimentations

Nous avons produit deux ensembles d'annotations qui diffèrent au niveau de la sélection des ressources et des patrons d'extraction. Le premier ensemble a utilisé une première version d'expressions régulières pour l'annotation de l'étiquette `mealTime` et une première collection de ressources terminologiques pour toutes les étiquettes référencées dans le tableau 5.3 à l'exception de `foodSupplement`. Le second ensemble a été produit après l'amélioration des patrons d'extraction pour `mealTime`, l'ajout d'une liste de termes pour des entités de type `foodSupplement` et l'ajout de ressources complémentaires pour les étiquettes `food`, `food2` et `drug`. Les ressources de MedDRA ont été

utilisées pour l'annotation de `sideeffect` dans le premier ensemble et pour l'annotation de `treatedDisease` dans le second ensemble.

2.2 Résultats

Le tableau 5.1 présente la macro-précision, le macro-rappel et la F-mesure des deux ensembles d'annotations effectuées avec OGMIOS. Le tableau 5.2 présente les mesures de précision, rappel et F-mesure les plus basses et les plus élevées dans chaque ensemble sans prendre en compte l'étiquette `foodSupplement`. Le tableau 5.3 présente, quant à lui, les mesures d'évaluation détaillées pour chaque étiquette annotée. Les valeurs les plus élevées entre le premier et le deuxième ensemble d'annotation ont été mises en gras.

TABLEAU 5.1 – Annotation des entités – Synthèse des résultats

	P_M	R_M	F_1
1er set	0.2103	**0.1793**	**0.1448**
2e set	**0.2468**	0.1313	0.1394

TABLEAU 5.2 – Annotation des entités – Comparatif (sans `foodSupp.`)

	P		R		F_1	
	min.	max.	min.	max.	min.	max.
1er set	0.0214	0.4010	0.0049	**0.4490**	0.0091	**0.3152**
2e set	**0.0417**	**0.6155**	**0.0056**	0.3806	**0.0099**	0.3046

Tableau 5.3 – Annotation des entités – Détails des résultats

	P		R		F_1	
	1er set	2e set	1er set	2e set	1er set	2e set
dosage	0.2295	0.2295	0.0815	0.0815	0.1203	0.1203
drug	0.3585	**0.4131**	**0.2812**	0.2413	**0.3152**	0.3046
duration	0.2558	0.2558	0.0210	0.0210	0.0388	0.0388
food	0.1592	**0.1924**	0.3673	**0.3806**	0.2222	**0.2556**
food2	0.0390	**0.0649**	0.0147	**0.0150**	0.0214	**0.0244**
food3	**0.0556**	0.0417	0.0049	**0.0056**	0.0091	**0.0099**
foodSupplement	-	**0.0054**	-	**0.0182**	-	**0.0083**
frequency	0.2908	0.2908	0.0628	0.0628	0.1033	0.1033
mealTime	0.0214	**0.0993**	**0.4490**	0.1910	0.0409	**0.1307**
modeadm	0.2263	0.2263	0.2023	0.2023	0.2137	0.2137
treatedDisease	0.2760	**0.6155**	**0.2300**	0.0498	**0.2509**	0.0920
sideeffect	**0.4010**	0.3154	**0.2574**	0.1933	**0.2574**	0.2397
moy. sans foodSupp.	0.2103	**0.2468**	**0.1793**	0.1313	**0.1448**	0.1394
moy. avec foodSupp.	-	0.2292	-	0.1219	-	0.1284

2.3 Discussion

Le tableau 5.1 montre que le deuxième ensemble obtient globalement une meilleure macro-précision que les annotations du premier set, tout en faisant baisser le macro-rappel et la F-mesure. Les détails du tableau 5.3 indiquent plus particulièrement que les résultats pour le deuxième ensemble d'annotation font baisser la précision pour l'étiquette `food3`, le rappel pour `mealTime` et le rappel et la F-mesure de `drug` et de `treatedDisease`. L'étiquette `sideeffect` obtient quant à elle de meilleurs résultats dans le premier set d'annotation ce qui s'explique par le fait qu'une partie des ressources ayant servi à son annotation dans le premier ensemble ont été utilisées dans le deuxième ensemble pour l'annotation de l'étiquette `treatedDisease`.

La constitution et le traitement des ressources pour *TermTagger* ainsi que l'écriture des partons d'extraction pour [Hamon and Grabar, 2010] ont pris du temps et bien que la précision augmente en même temps que le temps

passé à constituer ces ressources, cela ne permet pas d'augmenter le rappel des annotations par rapport à un corpus annoté manuellement. En effet, il reste difficile de répertorier toutes les entités recherchées. Les scores restent globalement trop bas pour que la méthode choisie puisse servir de base à une tâche d'extraction de relations.

3 Classification des relations

3.1 Expérimentations

Nous avons testé différentes configurations pour la conversion du contexte en un *dataset*. La relation entre les deux entités du contexte constituent les classes d'apprentissage et les étiquettes sémantiques des arguments de la relation constituent deux des attributs du *dataset*. Le contexte constitue donc le reste des attributs. Il est composé des mots apparaissant entre les deux arguments de la relation et ceux compris dans deux fenêtres fixées à 4 et 8 mots avant le premier argument et 4 et 8 mots après le second argument. Les mots du contexte sont représentés selon trois descripteurs différents :

— leurs formes fléchies

— leurs lemmes

— leurs étiquettes morpho-syntaxiques (*part-of-speech*)

Un exemple de contexte en fonction des descripteurs pour une fenêtre de 4 est donné dans le tableau 5.4. La valeur de ces attributs est numérique et dépend, dans un premier cas, de leur position dans le contexte par rapport aux arguments de la relation et, dans un second cas, de leur fréquence dans le contexte représentée par une valeur booléenne. Le tableau 5.5 présente le nombre d'attributs des *datasets* en fonction de la nature des attributs. Au total, 12 *datasets* ont été constitués en faisant varier la taille et les descripteurs du contexte et la valeur de leurs attributs.

Tableau 5.4 – Exemple fenêtre de 4

Formes fléchies	*POS*	Lemmes
clinical	JJ	clinical
use	NN	use
of	IN	of
oral	JJ	oral
oxymorphone	#	oxymorphone
IR	SYM	IR
and	CC	and
ER	WP$	ER
formulations	NNS	formulation
for	IN	for
the	DT	the
management	NN	management
of	IN	of
moderate	JJ	moderate
to	TO	to
severe	JJ	severe
pain	NN	pain
for	IN	for
different	JJ	different
types	NNS	type
of	IN	of

Tableau 5.5 – Nombre d'attributs en fonction des fenêtres

Attributs	fenêtre -4 +4	fenêtre -8 +8
Formes fléchies	3799	4505
Lemmes	3376	3915
Part-of-Speech	48	48

3.2 Résultats

L'évaluation de chaque modèle créé a été obtenue par validation croisée sur une base de 10 plis. Les tableaux 5.6 et 5.7 présentent respectivement les résultats pour la fenêtre incluant 4 tokens avant le premier argument et 4 tokens après le second argument et pour la fenêtre de 8 tokens avant le premier argument et 8 tokens après le second argument. Les scores les plus élevés ont été mis en gras et les détails de l'évaluation pour le modèle le plus performant sont précisés dans les tableaux 5.8 et 5.9. Dans ces deux derniers tableaux, les scores dépassant les 0.8 ont été mis en gras. La figure 5.1 représente un extrait du modèle entraîné sur le *dataset Lemmes, fréquence, fenêtre 8*.

TABLEAU 5.6 – Évaluation des modèles avec la fenêtre -4 +4 – validation croisée

	Formes fléchies		Lemmes		*POS*	
	position	fréquence	position	fréquence	position	fréquence
P_M	0.737	0.735	0.678	**0.752**	0.7	0.714
R_M	0.76	0.759	0.737	**0.776**	0.721	0.738
F_{1M}	0.743	0.742	0.69	**0.758**	0.708	0.723
Couverture des cas	87.1279 %	87.2084 %	**93.6042 %**	88.4553%	82.7031%	85.3178%
Classification incorrecte	23.9743 %	24.0949 %	26.3475 %	**22.4055 %**	27.8761 %	26.2269 %
Nombre de feuilles	308	297	**178**	277	321	326
Nb de nœuds	487	465	**227**	441	513	523
Temps de construction	129.46 sec	131.36 sec	12.73 sec	119.81 sec	0.78 sec	**0.75 sec**

TABLEAU 5.7 – Évaluation des modèles obtenus avec la fenêtre -8 +8 – validation croisée

	Formes fléchies		Lemmes		*POS*	
	position	fréquence	position	fréquence	position	fréquence
P_M	0.731	0.734	0.661	**0.748**	0.699	0.704
R_M	0.757	0.757	0.731	**0.771**	0.718	0.728
F_{1M}	0.74	0.741	0.686	**0.755**	0.707	0.715
Couverture des cas	86.3234 %	86.5245 %	**92.7997 %**	87.4497 %	82.0595 %	81.5768 %
Classification incorrecte	24.2558 %	24.2961 %	26.9107 %	**22.8882 %**	28.1577 %	27.1521 %
Nombre de feuilles	286	272	**184**	276	311	289
Nb de nœuds	459	431	**239**	439	509	465
Temps de construction	184.13 sec	150.32 sec	16.37 sec	146.59 sec	**1.76 sec**	1.78 sec

Tableau 5.8 – Détails par catégories pour le modèle *Lemmes, fréquence, fenêtre -4 +4* – validation croisée

Relations	P	R	$F1$
after	0	0	0
before	0.167	0.2	0.182
decrease_absorption	0.246	0.234	0.24
during	0.429	0.3	0.353
has_effect	**0.89**	**0.97**	**0.928**
has_sideeffect	**0.963**	**0.961**	**0.962**
improve_drug_effect	0	0	0
increase_absorption	0.265	0.173	0.209
increase_side_effect	0.678	**0.9**	0.773
negative_effect_on_drug	0.517	0.33	0.403
new_side_effect	0	0	0
no_effect_on_drug	0.655	0.641	0.648
positive_effect_on_drug	0.375	0.13	0.194
reduce_side_effect	0.8	0.174	0.286
relation	0.721	0.782	0.75
slow_absorption	0.529	0.429	0.474
slow_elimination	0.429	0.167	0.24
speed_elimination	0	0	0
speed_up_absorption	0	0	0
treat	**1**	**0.997**	**0.999**
without_food	0	0	0
worsen_drug_effect	0	0	0

TABLEAU 5.9 – Détails par catégories pour le modèle *Lemmes, fréquence, fenêtre -8 +8* – validation croisée

Relations	P	R	$F1$
after	0.286	0.182	0.222
before	0.25	0.2	0.222
decrease_absorption	0.353	0.281	0.313
during	0.4	0.2	0.267
has_effect	**0.89**	**0.974**	**0.93**
has_sideeffect	**0.965**	**0.961**	**0.963**
improve_drug_effect	0.077	0.071	0.074
increase_absorption	0.364	0.308	0.333
increase_side_effect	0.696	**0.87**	0.773
negative_effect_on_drug	0.483	0.308	0.376
new_side_effect	0	0	0
no_effect_on_drug	0.551	0.559	0.555
positive_effect_on_drug	0.25	0.087	0.129
reduce_side_effect	0.333	0.13	0.188
relation	0.714	0.78	0.746
slow_absorption	0.375	0.143	0.207
slow_elimination	0.5	0.278	0.357
speed_elimination	0	0	0
speed_up_absorption	0	0	0
treat	**1**	**0.997**	**0.999**
without_food	0.222	0.087	0.125
worsen_drug_effect	0.2	0.063	0.095

Figure 5.1 – Extrait modèle *Lemmes, fréquence, fenêtre +8 -8*

```
argument2 = drugEffect
|   argument1 = food
|   |   voc_d-amphetamine <= 0
|   |   |   voc_) <= 0
|   |   |   |   voc_. <= 0
|   |   |   |   |   voc_juice <= 0: relation (5.0/2.0)
|   |   |   |   |   voc_juice > 0: improve_drug_effect (3.0)
|   |   |   |   voc_. > 0
|   |   |   |   |   voc_drug <= 1: negative_effect_on_drug (5.0/2.0)
|   |   |   |   |   voc_drug > 1: relation (2.0)
|   |   |   voc_) > 0: increase_absorption (2.0/1.0)
|   |   voc_d-amphetamine > 0: no_effect_on_drug (4.0/1.0)
|   argument1 = food2: has_effect (0.0)
|   argument1 = food3: has_effect (0.0)
|   argument1 = food4: has_effect (0.0)
|   argument1 = drug: has_effect (156.0)
|   argument1 = drug2: has_effect (0.0)
|   argument1 = treatedDisease: has_effect (0.0)
|   argument1 = sideeffect: has_effect (0.0)
|   argument1 = mealTime: has_effect (0.0)
|   argument1 = drugEffect: has_effect (0.0)
|   argument1 = foodSupplement: positive_effect_on_drug (2.0)
|   argument1 = numbers: has_effect (0.0)
|   argument1 = modeadm: has_effect (0.0)
|   argument1 = frequency: has_effect (0.0)
|   argument1 = dosage: has_effect (0.0)
|   argument1 = duration: has_effect (0.0)
|   argument1 = other: has_effect (0.0)
|   argument1 = ABS: has_effect (0.0)
argument2 = foodSupplement
|   voc_but <= 0
|   |   voc_lead <= 0
|   |   |   voc_exposure <= 0
|   |   |   |   argument1 = food: relation (0.0)
|   |   |   |   argument1 = food2: relation (1.0)
|   |   |   |   argument1 = food3: relation (0.0)
|   |   |   |   argument1 = food4: relation (0.0)
|   |   |   |   argument1 = drug
|   |   |   |   |   voc_specific <= 0
|   |   |   |   |   |   voc_occurrence <= 0: relation (14.0/2.0)
|   |   |   |   |   |   voc_occurrence > 0: without_food (2.0)
|   |   |   |   |   voc_specific > 0: without_food (2.0)
|   |   |   |   argument1 = drug2: relation (0.0)
|   |   |   |   argument1 = treatedDisease: relation (0.0)
|   |   |   |   argument1 = sideeffect
|   |   |   |   |   voc_any <= 0: has_sideeffect (11.0/1.0)
|   |   |   |   |   voc_any > 0: increase_side_effect (2.0/1.0)
|   |   |   |   argument1 = mealTime: relation (0.0)
|   |   |   |   argument1 = drugEffect: relation (0.0)
|   |   |   |   argument1 = foodSupplement: relation (0.0)
|   |   |   |   argument1 = numbers: relation (0.0)
|   |   |   |   argument1 = modeadm: relation (0.0)
|   |   |   |   argument1 = frequency: relation (0.0)
|   |   |   |   argument1 = dosage: relation (0.0)
|   |   |   |   argument1 = duration: relation (0.0)
|   |   |   |   argument1 = other: relation (0.0)
|   |   |   |   argument1 = ABS: increase_side_effect (2.0)
|   |   |   voc_exposure > 0: positive_effect_on_drug (2.0)
|   |   voc_lead > 0: increase_absorption (2.0/1.0)
|   voc_but > 0: worsen_drug_effect (2.0)
argument2 = numbers: relation (0.0)
argument2 = modeadm: relation (0.0)
```

3.3 Discussion

On note que l'agrandissement de la fenêtre rallonge le temps de calcul et diminue la taille des modèles sans pour autant améliorer leurs performances en terme de macro-précision et macro-rappel. Cependant lorsque l'on regarde les résultats du modèle le plus performant en détail, on observe que l'allongement de la fenêtre du contexte permet d'obtenir des résultats pour `after`, `without_food` et `worsen_drug_effect` alors que ces relations n'étaient pas identifiées avec une fenêtre de 4. D'autres relations observent aussi des augmentations dans leurs résultats.

Concernant les temps d'entraînement des modèles, les plus courts sont ceux entraînés à partir des étiquettes morpho-syntaxiques en descripteurs, le nombre d'attributs étant en effet beaucoup plus faibles que pour les lemmes ou les formes fléchies. Les arbres qui en résultent sont cependant plus volumineux que pour les autres *datasets*.

FIGURE 5.2 – Extrait du modèle *Formes fléchies, position, fenêtre +4 -4*

```
argument1 = drug
|   argument1 = food2: relation (9.0)
|   argument1 = food3: relation (9.0)
|   argument1 = food4: relation (1.0)
|   argument1 = drug: relation (0.0)
|   argument1 = drug2: relation (0.0)
|   argument1 = treatedDisease: treat (56.0)
|   argument1 = sideeffect: has_sideeffect (155.0)
(...)
|   argument1 = drugEffect: has_effect (63.0)
(...)
|   argument1 = numbers: relation (0.0)
|   argument1 = modeadm: relation (0.0)
|   argument1 = frequency: relation (0.0)
|   argument1 = dosage: relation (0.0)
|   argument1 = duration: relation (0.0)
|   argument1 = other: relation (0.0)
|   argument1 = ABS: has_effect (44.0/30.0)
```

Les modèles construits par J48 permettent dans certains cas d'attribuer une relation à un contexte uniquement en fonction de l'étiquette du premier et du deuxième argument de la relation. C'est le cas dans l' exemple en figure 5.2. Le patron d'extraction comporte dans ces cas de figure uniquement les étiquettes sémantiques des deux arguments de la relation.

Cela est dû notamment au fait que certaines relations du corpus ont été définies de telle manière qu'elles prennent une étiquette particulière en premier et-ou en deuxième argument. Ainsi, la relation de type `treat` ne prend simultanément en argument que les étiquettes `treatedDisease` et `drug`.

4 Conclusion

Nous avons constitué deux ensembles d'annotations qui ont été ensuite évalué avec BratEval. Le deuxième ensemble présente une meilleure macro-précision que le premier ensemble mais un macro-rappel et une F-mesure plus faible même si globalement les résultats par étiquettes sont plus élevés.

Nous avons expérimenté plusieurs configurations pour entraîner 12 modèles de classification de relations. Les *datasets* varient selon la taille de la fenêtre du contexte, les descripteurs des attributs (formes fléchies, lemmes ou étiquettes morpho-syntaxiques) et leur valeur (fréquence d'apparition ou position). On obtient des résultats légèrement meilleurs avec la configuration *Lemmes, fréquence, fenêtre 8.*

Chapitre 6

Conclusion et Perspectives

1 Annotations

1.1 Conclusion

Nous avons donc effectué deux ensembles d'annotations avec deux ensembles de ressources et de patrons d'extraction différents. Les ressources et patrons du premier ensemble ont été complétés pour le deuxième ensemble et les ressources issues de la base de données MedDRA ont été utilisées dans le premier ensemble pour `sideeffect` et dans le deuxième ensemble pour `treatedDisease`. L'évaluation des deux ensembles d'annotations montre que l'ajout de ressources terminologiques et de patrons d'extraction permet d'améliorer légèrement les scores obtenus pour chaque étiquette. Cependant la macro-évaluation est peu impactée par cet apport et on note même un recul du macro-rappel et de la F-mesure. Le problème du choix à effectuer entre deux étiquettes pour une seule entité se pose toujours et les résultats d'évaluation restent trop faibles pour constituer une base à un apprentissage automatique.

1.2 Perspectives

Les expérimentations en annotation d'entités par projection de ressources montrent qu'il reste nécessaire d'effectuer de l'apprentissage automatique pour annoter un corpus en entités biomédicales plus spécifiquement pour les entités dont on possède peu de ressources comme `drugEffect` ou celles dont l'anno-

tation dépend du contexte à l'exemple de `treatedDisease` et de `sideeffect`. Il serait intéressant d'adapter les outils de reconnaissance d'entités existants, comme ceux développés lors de challenge, pour permettre la détection des entités alimentaires tels que `food`, `food2`, `food3` et `foodSupplement`.

2 Extraction de relations

2.1 Conclusion

A défaut d'avoir procédé à une fouille itérative du corpus pour en extraire des patrons de relations comme dans les travaux de [Hearst, 1992], [Jean-Louis et al., 2011], [Riloff, 1996] ou de [Meng and Morioka, 2015], plusieurs modèles de classification des relations du corpus ont été construits en fonction de la taille du contexte et de la fréquence ou de la position d'apparition des formes fléchies, des lemmes ou des étiquettes morpho-syntaxiques dans le contexte. Les résultats d'évaluation varient peu en terme de macro-mesure mais l'allongement de la fenêtre permet d'augmenter le nombre de relations reconnues. Le trait fréquence des lemmes semble par ailleurs augmenter les scores d'évaluation.

2.2 Perspectives

Il serait intéressant d'évaluer les modèles sur un plus grand nombre de données test. Il semble par ailleurs possible de convertir les arbres obtenus en exploitant la position des éléments du contexte en des patrons d'extraction parmi lesquels il s'agirait ensuite de faire le tri. Il serait aussi intéressant de combiner les traits des différents ensembles (formes fléchies, lemmes et étiquettes morpho-syntaxiques) pour la classification des relations. Il serait donc nécessaire de continuer les expérimentations afin de déterminer les caractéristiques des traits qui permettraient d'obtenir un ensemble optimisé pour l'apprentissage de patrons d'extraction. Par ailleurs, l'enrichissement sémantique du contexte avec les ressources terminologiques permettrait peut-être d'obtenir des patrons d'extraction basés sur des caractéristiques sémantiques ce qui permettrait d'effectuer un apprentissage sur des données annotées sémantiquement.

Annexes

Liste des étiquettes POS :

SENT . de fin de phrase

DT déterminant

NNS nom au pluriel

IN préposition ou conjonction de subordination

JJ adjectif

NN nom au singulier ou massif

WRB WH-adverbe

VBG gérondif ou participe présent

SYM symbole

LS marqueur d'objet liste

VBN participe passé

VBP verbe au présent autre qu'à la 3ème personne du singulier

TO to

VB base verbale

(ponctuation

) ponctuation

CC conjonction de coordination

VBD verbe au passé

” ponctuation

ponctuation

CD nombre cardinal

JJS adjectif superlatif

MD modal

RB adverbe

PP pronom personnel

EX 'there' existentiel

VBZ verbe au présent de la 3ème personne du singulier

PP$ pronom possessif

JJR adjectif comparatif

NP nom propre

FW mot étranger

UH interjection

RP particule

WDT WH-déterminant

“ ponctuation

WP WH-pronom

NPS nom propre au pluriel

: ponctuation

RBS adverbe superlatif

RBR adverbe comparatif

WP$ wh-pronom possessif

POS possessif ’s

$ ponctuation

PDT prédéterminant (ex : ’all both’)

ponctuation

`https://ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

FIGURE 6.1 – Détails par catégories pour le modèle *Formes fléchies, position, fenêtre -4 +4*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0.091     0.001      0.25      0.091     0.133      0.626     0.082    after
                 0         0.002      0         0         0          0.597     0.027    before
                 0.297     0.019      0.288     0.297     0.292      0.769     0.185    decrease_absorption
                 0.3       0.002      0.375     0.3       0.333      0.895     0.32     during
                 0.974     0.012      0.89      0.974     0.93       0.999     0.994    has_effect
                 0.961     0.006      0.97      0.961     0.965      0.994     0.988    has_sideeffect
                 0.143     0.004      0.167     0.143     0.154      0.661     0.045    improve_drug_effect
                 0.154     0.008      0.286     0.154     0.2        0.677     0.094    increase_absorption
                 0.879     0.047      0.665     0.879     0.757      0.92      0.609    increase_side_effect
                 0.308     0.021      0.354     0.308     0.329      0.827     0.275    negative_effect_on_drug
                 0         0          0         0         0          0.452     0.002    new_side_effect
                 0.524     0.026      0.551     0.524     0.537      0.837     0.434    no_effect_on_drug
                 0.087     0.001      0.4       0.087     0.143      0.843     0.135    positive_effect_on_drug
                 0.261     0          0.857     0.261     0.4        0.815     0.305    reduce_side_effect
                 0.761     0.125      0.708     0.761     0.733      0.876     0.679    relation
                 0         0.003      0         0         0          0.624     0.016    slow_absorption
                 0.222     0.002      0.444     0.222     0.296      0.704     0.162    slow_elimination
                 0         0          0         0         0          ?         ?        speed_elimination
                 0         0          0         0         0          0.558     0.003    speed_up_absorption
                 0.997     0          1         0.997     0.999      0.999     0.998    treat
                 0.043     0.002      0.167     0.043     0.069      0.668     0.07     without_food
                 0         0.001      0         0         0          0.649     0.106    worsen_drug_effect
Weighted Avg.    0.76      0.045      0.737     0.76      0.743      0.907     0.708
```

Figure 6.2 – Détails par catégories pour le modèle *Formes fléchies, fréquence, fenêtre -4 +4*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0         0.003      0         0         0          0.579     0.02     after
                 0         0.004      0         0         0          0.596     0.022    before
                 0.234     0.017      0.268     0.234     0.25       0.733     0.176    decrease_absorption
                 0.2       0.002      0.286     0.2       0.235      0.743     0.187    during
                 0.974     0.012      0.89      0.974     0.93       0.999     0.994    has_effect
                 0.961     0.006      0.97      0.961     0.965      0.994     0.988    has_sideeffect
                 0.143     0.005      0.133     0.143     0.138      0.656     0.039    improve_drug_effect
                 0.231     0.011      0.308     0.231     0.264      0.652     0.092    increase_absorption
                 0.87      0.048      0.658     0.87      0.75       0.913     0.612    increase_side_effect
                 0.275     0.015      0.417     0.275     0.331      0.822     0.286    negative_effect_on_drug
                 0         0          0         0         0          0.44      0.002    new_side_effect
                 0.579     0.022      0.618     0.579     0.598      0.846     0.481    no_effect_on_drug
                 0.087     0.003      0.222     0.087     0.125      0.81      0.148    positive_effect_on_drug
                 0.261     0.001      0.75      0.261     0.387      0.814     0.263    reduce_side_effect
                 0.755     0.13       0.697     0.755     0.725      0.874     0.669    relation
                 0.048     0.002      0.143     0.048     0.071      0.605     0.03     slow_absorption
                 0.167     0.002      0.375     0.167     0.231      0.72      0.148    slow_elimination
                 0         0          0         0         0          ?         ?        speed_elimination
                 0         0          0         0         0          0.521     0.002    speed_up_absorption
                 0.997     0          1         0.997     0.999      0.999     0.998    treat
                 0.043     0.002      0.143     0.043     0.067      0.712     0.092    without_food
                 0         0.001      0         0         0          0.707     0.187    worsen_drug_effect
Weighted Avg.    0.759     0.047      0.735     0.759     0.742      0.904     0.708
```

Figure 6.3 – Détails par catégories pour le modèle *Lemmes, position, fenêtre -4 +4*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0.091     0.001      0.25      0.091     0.133      0.799     0.111    after
                 0         0.001      0         0         0          0.781     0.02     before
                 0.063     0.006      0.211     0.063     0.096      0.755     0.08     decrease_absorption
                 0         0          0         0         0          0.833     0.101    during
                 0.974     0.012      0.89      0.974     0.93       0.999     0.994    has_effect
                 0.965     0.007      0.968     0.965     0.967      0.998     0.991    has_sideeffect
                 0.143     0.001      0.5       0.143     0.222      0.732     0.207    improve_drug_effect
                 0.038     0.006      0.118     0.038     0.058      0.665     0.043    increase_absorption
                 0.975     0.056      0.651     0.975     0.781      0.963     0.681    increase_side_effect
                 0.077     0.005      0.35      0.077     0.126      0.82      0.154    negative_effect_on_drug
                 0         0          0         0         0          0.373     0.002    new_side_effect
                 0.152     0.022      0.297     0.152     0.201      0.834     0.192    no_effect_on_drug
                 0.087     0          1         0.087     0.16       0.782     0.121    positive_effect_on_drug
                 0         0.001      0         0         0          0.818     0.051    reduce_side_effect
                 0.797     0.208      0.603     0.797     0.687      0.866     0.646    relation
                 0         0.001      0         0         0          0.703     0.022    slow_absorption
                 0         0.001      0         0         0          0.838     0.126    slow_elimination
                 0         0          0         0         0          ?         ?        speed_elimination
                 0         0          0         0         0          0.397     0.002    speed_up_absorption
                 0.997     0          1         0.997     0.999      0.999     0.998    treat
                 0         0.002      0         0         0          0.817     0.103    without_food
                 0         0.001      0         0         0          0.783     0.14     worsen_drug_effect
Weighted Avg.    0.737     0.069      0.678     0.737     0.69       0.912     0.682
```

FIGURE 6.4 – Détails par catégories pour le modèle *POS, position, fenêtre -4 +4*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0.182     0.005      0.133     0.182      0.154      0.626     0.04     after
                 0         0.001      0         0          0          0.697     0.161    before
                 0.109     0.02       0.127     0.109      0.118      0.645     0.056    decrease_absorption
                 0.4       0.004      0.286     0.4        0.333      0.793     0.197    during
                 0.974     0.012      0.89      0.974      0.93       0.999     0.994    has_effect
                 0.965     0.006      0.97      0.965      0.968      0.995     0.985    has_sideeffect
                 0.071     0.004      0.091     0.071      0.08       0.506     0.029    improve_drug_effect
                 0.135     0.015      0.163     0.135      0.147      0.566     0.045    increase_absorption
                 0.728     0.035      0.688     0.728      0.707      0.912     0.624    increase_side_effect
                 0.264     0.014      0.414     0.264      0.322      0.739     0.251    negative_effect_on_drug
                 0         0.001      0         0          0          0.623     0.051    new_side_effect
                 0.4       0.042      0.372     0.4        0.385      0.75      0.238    no_effect_on_drug
                 0.087     0.005      0.143     0.087      0.108      0.773     0.079    positive_effect_on_drug
                 0.174     0.002      0.4       0.174      0.242      0.702     0.191    reduce_side_effect
                 0.72      0.158      0.644     0.72       0.68       0.828     0.595    relation
                 0.095     0.002      0.25      0.095      0.138      0.613     0.037    slow_absorption
                 0.278     0.001      0.625     0.278      0.385      0.734     0.225    slow_elimination
                 0         0          0         0          0          ?         ?        speed_elimination
                 0         0.001      0         0          0          0.496     0.002    speed_up_absorption
                 0.997     0          1         0.997      0.999      0.999     0.998    treat
                 0         0.002      0         0          0          0.633     0.063    without_food
                 0         0.001      0         0          0          0.52      0.019    worsen_drug_effect
Weighted Avg.    0.721     0.054      0.7       0.721      0.708      0.875     0.667
```

FIGURE 6.5 – Détails par catégories pour le modèle *POS, fréquence, fenêtre -4 +4*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0.091     0.004      0.091     0.091      0.091      0.534     0.014    after
                 0         0.002      0         0          0          0.597     0.03     before
                 0.234     0.015      0.294     0.234      0.261      0.684     0.117    decrease_absorption
                 0.3       0.002      0.333     0.3        0.316      0.744     0.26     during
                 0.974     0.012      0.89      0.974      0.93       0.999     0.994    has_effect
                 0.965     0.006      0.97      0.965      0.968      0.997     0.992    has_sideeffect
                 0.143     0.004      0.154     0.143      0.148      0.589     0.049    improve_drug_effect
                 0.115     0.015      0.14      0.115      0.126      0.571     0.048    increase_absorption
                 0.766     0.036      0.693     0.766      0.728      0.943     0.67     increase_side_effect
                 0.198     0.018      0.295     0.198      0.237      0.78      0.23     negative_effect_on_drug
                 0.25      0.001      0.25      0.25       0.25       0.73      0.083    new_side_effect
                 0.448     0.039      0.414     0.448      0.43       0.787     0.324    no_effect_on_drug
                 0.087     0.003      0.222     0.087      0.125      0.675     0.066    positive_effect_on_drug
                 0.13      0.003      0.3       0.13       0.182      0.718     0.14     reduce_side_effect
                 0.755     0.138      0.684     0.755      0.718      0.862     0.666    relation
                 0.048     0.003      0.125     0.048      0.069      0.606     0.034    slow_absorption
                 0.278     0.002      0.556     0.278      0.37       0.763     0.208    slow_elimination
                 0         0          0         0          0          ?         ?        speed_elimination
                 0         0.001      0         0          0          0.493     0.002    speed_up_absorption
                 0.997     0          1         0.997      0.999      0.999     0.998    treat
                 0         0.003      0         0          0          0.599     0.028    without_food
                 0.063     0.002      0.143     0.063      0.087      0.602     0.052    worsen_drug_effect
Weighted Avg.    0.738     0.049      0.714     0.738      0.723      0.893     0.697
```

Figure 6.6 – Détails par catégories pour le modèle *Formes fléchies, position, fenêtre -8 +8*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0.182     0.002      0.25      0.182     0.211      0.581     0.049    after
                 0         0.002      0         0         0          0.497     0.002    before
                 0.188     0.018      0.218     0.188     0.202      0.709     0.141    decrease_absorption
                 0.4       0.003      0.333     0.4       0.364      0.743     0.247    during
                 0.974     0.012      0.89      0.974     0.93       0.999     0.994    has_effect
                 0.961     0.007      0.965     0.961     0.963      0.993     0.984    has_sideeffect
                 0         0.005      0         0         0          0.616     0.022    improve_drug_effect
                 0.269     0.012      0.333     0.269     0.298      0.705     0.157    increase_absorption
                 0.887     0.042      0.693     0.887     0.778      0.927     0.622    increase_side_effect
                 0.352     0.012      0.525     0.352     0.421      0.828     0.335    negative_effect_on_drug
                 0         0          0         0         0          0.608     0.008    new_side_effect
                 0.434     0.033      0.45      0.434     0.442      0.835     0.337    no_effect_on_drug
                 0.087     0.003      0.2       0.087     0.121      0.836     0.116    positive_effect_on_drug
                 0.174     0.001      0.667     0.174     0.276      0.771     0.156    reduce_side_effect
                 0.766     0.125      0.708     0.766     0.736      0.87      0.692    relation
                 0         0.002      0         0         0          0.705     0.031    slow_absorption
                 0.222     0.003      0.333     0.222     0.267      0.762     0.13     slow_elimination
                 0         0          0         0         0          ?         ?        speed_elimination
                 0         0          0         0         0          0.469     0.002    speed_up_absorption
                 0.997     0          1         0.997     0.999      0.999     0.998    treat
                 0         0.003      0         0         0          0.721     0.054    without_food
                 0         0.002      0         0         0          0.678     0.14     worsen_drug_effect
Weighted Avg.    0.757     0.045      0.731     0.757     0.74       0.905     0.707
```

Figure 6.7 – Détails par catégories pour le modèle *Formes fléchies, fréquence, fenêtre -8 +8*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0.091     0.003      0.125     0.091     0.105      0.67      0.046    after
                 0         0.002      0         0         0          0.496     0.002    before
                 0.219     0.017      0.255     0.219     0.235      0.72      0.172    decrease_absorption
                 0.3       0.002      0.333     0.3       0.316      0.742     0.295    during
                 0.974     0.012      0.89      0.974     0.93       0.999     0.994    has_effect
                 0.963     0.006      0.97      0.963     0.966      0.996     0.99     has_sideeffect
                 0.143     0.004      0.182     0.143     0.16       0.616     0.054    improve_drug_effect
                 0.308     0.008      0.457     0.308     0.368      0.678     0.187    increase_absorption
                 0.849     0.043      0.677     0.849     0.753      0.929     0.661    increase_side_effect
                 0.286     0.015      0.419     0.286     0.34       0.815     0.27     negative_effect_on_drug
                 0         0          0         0         0          0.414     0.002    new_side_effect
                 0.545     0.029      0.537     0.545     0.541      0.862     0.389    no_effect_on_drug
                 0.087     0.002      0.333     0.087     0.138      0.856     0.151    positive_effect_on_drug
                 0.217     0.003      0.417     0.217     0.286      0.773     0.157    reduce_side_effect
                 0.751     0.137      0.685     0.751     0.716      0.857     0.654    relation
                 0.048     0.003      0.125     0.048     0.069      0.671     0.045    slow_absorption
                 0.222     0.002      0.444     0.222     0.296      0.702     0.153    slow_elimination
                 0         0          0         0         0          ?         ?        speed_elimination
                 0         0          0         0         0          0.418     0.002    speed_up_absorption
                 0.997     0          1         0.997     0.999      0.999     0.998    treat
                 0.043     0.002      0.167     0.043     0.069      0.614     0.045    without_food
                 0.063     0.001      0.333     0.063     0.105      0.714     0.119    worsen_drug_effect
Weighted Avg.    0.757     0.048      0.734     0.757     0.741      0.901     0.704
```

FIGURE 6.8 – Détails par catégories pour le modèle *Lemmes, position, fenêtre -8 +8*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0.273     0.002      0.333     0.273      0.3        0.714     0.225    after
                 0         0.002      0         0          0          0.684     0.011    before
                 0.078     0.007      0.238     0.078      0.118      0.733     0.086    decrease_absorption
                 0.1       0.002      0.2       0.1        0.133      0.837     0.12     during
                 0.974     0.012      0.89      0.974      0.93       0.999     0.994    has_effect
                 0.965     0.008      0.963     0.965      0.964      0.999     0.995    has_sideeffect
                 0.143     0.002      0.286     0.143      0.19       0.666     0.071    improve_drug_effect
                 0.019     0.008      0.048     0.019      0.027      0.668     0.056    increase_absorption
                 0.975     0.054      0.656     0.975      0.785      0.965     0.682    increase_side_effect
                 0         0.006      0         0          0          0.816     0.109    negative_effect_on_drug
                 0         0          0         0          0          0.677     0.008    new_side_effect
                 0.166     0.023      0.312     0.166      0.216      0.815     0.204    no_effect_on_drug
                 0.087     0.001      0.5       0.087      0.148      0.781     0.122    positive_effect_on_drug
                 0         0.001      0         0          0          0.838     0.045    reduce_side_effect
                 0.78      0.202      0.605     0.78       0.682      0.872     0.668    relation
                 0         0          0         0          0          0.678     0.014    slow_absorption
                 0         0.002      0         0          0          0.754     0.026    slow_elimination
                 0         0          0         0          0          ?         ?        speed_elimination
                 0         0          0         0          0          0.666     0.005    speed_up_absorption
                 0.997     0          1         0.997      0.999      0.999     0.998    treat
                 0         0.003      0         0          0          0.744     0.067    without_food
                 0         0.002      0         0          0          0.601     0.033    worsen_drug_effect
Weighted Avg.    0.731     0.067      0.661     0.731      0.686      0.91      0.686
```

FIGURE 6.9 – Détails par catégories pour le modèle *POS, position, fenêtre -8 +8*

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  PRC Area  Class
                 0.091     0.004      0.083     0.091      0.087      0.584     0.027    after
                 0         0.003      0         0          0          0.495     0.002    before
                 0.141     0.018      0.17      0.141      0.154      0.653     0.081    decrease_absorption
                 0.2       0.003      0.222     0.2        0.211      0.641     0.093    during
                 0.974     0.012      0.89      0.974      0.93       0.999     0.994    has_effect
                 0.963     0.007      0.968     0.963      0.965      0.994     0.981    has_sideeffect
                 0         0.003      0         0          0          0.467     0.005    improve_drug_effect
                 0.115     0.017      0.125     0.115      0.12       0.536     0.053    increase_absorption
                 0.728     0.034      0.693     0.728      0.71       0.93      0.654    increase_side_effect
                 0.308     0.02       0.368     0.308      0.335      0.74      0.242    negative_effect_on_drug
                 0         0.002      0         0          0          0.495     0.002    new_side_effect
                 0.366     0.038      0.371     0.366      0.368      0.739     0.235    no_effect_on_drug
                 0.13      0.005      0.188     0.13       0.154      0.617     0.086    positive_effect_on_drug
                 0.174     0.004      0.286     0.174      0.216      0.615     0.089    reduce_side_effect
                 0.717     0.146      0.661     0.717      0.688      0.834     0.598    relation
                 0.048     0.002      0.2       0.048      0.077      0.598     0.035    slow_absorption
                 0.278     0.006      0.263     0.278      0.27       0.679     0.187    slow_elimination
                 0         0          0         0          0          ?         ?        speed_elimination
                 0         0.001      0         0          0          0.622     0.051    speed_up_absorption
                 0.997     0          1         0.997      0.999      0.999     0.998    treat
                 0         0.003      0         0          0          0.549     0.021    without_food
                 0         0.004      0         0          0          0.576     0.031    worsen_drug_effect
Weighted Avg.    0.718     0.051      0.699     0.718      0.707      0.873     0.668
```

Figure 6.10 – Détails par catégories pour le modèle *POS, fréquence, fenêtre -8 +8*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area	Class
	0	0.004	0	0	0	0.489	0.004	after
	0	0.002	0	0	0	0.497	0.002	before
	0.078	0.02	0.093	0.078	0.085	0.607	0.052	decrease_absorption
	0.6	0.002	0.545	0.6	0.571	0.896	0.506	during
	0.974	0.012	0.89	0.974	0.93	0.999	0.994	has_effect
	0.961	0.009	0.959	0.961	0.96	0.993	0.983	has_sideeffect
	0.143	0.004	0.182	0.143	0.16	0.518	0.032	improve_drug_effect
	0.115	0.017	0.128	0.115	0.121	0.566	0.056	increase_absorption
	0.782	0.032	0.722	0.782	0.751	0.934	0.689	increase_side_effect
	0.242	0.021	0.301	0.242	0.268	0.704	0.151	negative_effect_on_drug
	0.5	0.002	0.286	0.5	0.364	0.747	0.126	new_side_effect
	0.352	0.038	0.367	0.352	0.359	0.697	0.235	no_effect_on_drug
	0.13	0.006	0.176	0.13	0.15	0.625	0.068	positive_effect_on_drug
	0.217	0.004	0.333	0.217	0.263	0.713	0.113	reduce_side_effect
	0.742	0.137	0.682	0.742	0.711	0.839	0.632	relation
	0	0.002	0	0	0	0.625	0.036	slow_absorption
	0.111	0.003	0.2	0.111	0.143	0.532	0.04	slow_elimination
	0	0	0	0	0	?	?	speed_elimination
	0	0	0	0	0	0.594	0.043	speed_up_absorption
	0.997	0	1	0.997	0.999	0.999	0.998	treat
	0.13	0.004	0.25	0.13	0.171	0.608	0.055	without_food
	0	0.002	0	0	0	0.546	0.011	worsen_drug_effect
Weighted Avg.	0.728	0.049	0.704	0.728	0.715	0.872	0.678	

Bibliographie

[Abacha et al., 2015] Abacha, A. B., Chowdhury, M. F. M., Karanasiou, A., Mrabet, Y., Lavelli, A., and Zweigenbaum, P. (2015). Text mining for pharmacovigilance : Using machine learning for drug name recognition and drug–drug interaction extraction and classification. *Journal of Biomedical Informatics*, 58 :122–132.

[Aramaki et al., 2010] Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Masuichi, H., Waki, K., and Ohe, K. (2010). Extraction of adverse drug effects from clinical records. 160 :739–43.

[Aussenac-Gilles and Condamines, 2009] Aussenac-Gilles, N. and Condamines, A. (2009). Variation syntaxique et contextuelle dans la mise au point de patrons de relations sémantiques. *Filtrage sémantique, Hermes/Lavoisier*, pages 115–149.

[Bach and Badaskar, 2007] Bach, N. and Badaskar, S. (2007). A review of relation extraction.

[Cellier and Charnois, 2010] Cellier, P. and Charnois, T. (2010). Fouille de données séquentielle d'itemsets pour l'apprentissage de patrons linguistiques. In *17e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010), Jul 2010, Montréal, Canada. 6 p.*

[Charnois et al., 2009] Charnois, T., Plantevit, M., Rigotti, C., and Crémilleux, B. (2009). Fouille de données séquentielles pour l'extraction d'information dans les textes. *Traitement Automatique des Langues, ATALA*, pages 59–87.

[Cohen and Hunter, 2008] Cohen, K. B. and Hunter, L. (2008). Getting started in text mining. *PLoS Comput Biol*, 4.

[Frank et al., 2016] Frank, E., Hall, M. A., and Witten, I. H. (2016). *The WEKA Workbench*, online appendix for "data mining : practical machine learning tools and techniques" edition.

[Fundel et al., 2007] Fundel, K., Küffner, R., and Zimmer, R. (2007). Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3) :365–371.

[Grabar et al., 2015] Grabar, N., Amiot, D., Mougin, F., Thiessard, F., and Hamon, T. (2015). Rapport final projet meshs Émergent pomelo : Patologies medicaments alimentation.

[Hamon and Grabar, 2010] Hamon, T. and Grabar, N. (2010). Linguistic approach for identification of medication names and related information in clinical narratives. *Journal of American Medical Informatics Association*, 17(5) :549–554. PMID : 20819862.

[Hamon and Nazarenko, 2008] Hamon, T. and Nazarenko, A. (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience. 49(2) :127–154.

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of COLING-92*, page 539 :545.

[Huang et al., 2004] Huang, M., Zhu, X., Payan, D. G., Qu, K., and Li, M. (2004). Discovering patterns to extract protein-protein interactions from full biomedical texts. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, page 22 :28, Geneva, Switzerland.

[Jean-Louis et al., 2011] Jean-Louis, L., Besançon, R., Ferret, O., and Durand, A. (2011). Une approche faiblement supervisée pour l'extraction de relations à large échelle. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier, France. Association pour le Traitement Automatique des Langues.

[Jiang et al., 2016] Jiang, X., Wang, Q., Li, P., and Wang, B. (2016). Relation extraction with multi-instance multi-label convolutional neural networks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, pages 1471–1480.

[Jovanovik et al., 2015] Jovanovik, M., Bogojeska, A., Trajanov, D., and Kocarev, L. (2015). Inferring cuisine - drug interactions using the linked data approach. *Scientific Reports*.

[Liu et al., 2016] Liu, S., Tang, B., Chen, Q., and Wang, X. (2016). Drug-drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, 2016.

[Meng and Morioka, 2015] Meng, F. and Morioka, C. (2015). Automating the generation of lexical patterns for processing free text in clinical documents. *JAMIA*, 22 :980–986.

[Morin, 1999] Morin, E. (1999). *Extraction de liens semantiques entre termes a partir de corpus de textes techniques*. PhD thesis. Thèse de doctorat dirigée par Jacquemin, Christian Sciences et techniques. Informatique Nantes 1999.

[Nguyen and Grishman, 2015] Nguyen, T. H. and Grishman, R. (2015). Event detection and domain adaptation with convolutional neural networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 365–371.

[Riloff, 1996] Riloff, E. (1996). Automatically generating extraction patterns from untagged text. *AAAI-96 Proceedings*.

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*.

[Sebastiani, 2001] Sebastiani, F. (2001). Machine learning in automated text categorization. *CoRR*, cs.IR/0110053.

[T. et al., 2009] T., B., J., W., R., C., H.-W., M., and V., S. (2009). Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS ONE*, 4.

[Tatonetti et al., 2012] Tatonetti, N. P., Fernald, G. H., and Altman, R. B. (2012). A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *JAMIA*, 19 :79 :85.

[Wei et al., 2016] Wei, C.-H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., Wiegers, T. C., and Lu, Z. (2016). Assessing the state of the art in biomedical relation extraction : overview of the biocreative v chemical-disease relation (cdr) task. *Database (2016) Vol. 2016 : article ID ; doi :*, 2016 :1–8.

[Zelenko et al., 2003] Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research*, pages 1083–1106.

[Zhou et al., 2014] Zhou, D., Zhong, D., and He, Y. (2014). Biomedical relation extraction : From binary to complex. *Computational and Mathematical Methods in Medicine*, 2014 :18.