Annotations d'entités et de relations sur des résumés d'articles scientifiques pour la détection d'interactions entre aliments et médicaments

Tsanta Randriatsitohaina, Cyril Grouin, Pierrick Bedouch, Georgeta Bordea, Amélie Daveluy, Vincent Depras, Natalia Grabar, Ghada Miremont-Salamé, Fleur Mougin, Cécile Pageot, et al.

Annotations d'entités et de relations sur des résumés d'articles scientifiques pour la détection d'interactions entre aliments et médicaments

Tsanta Randriatsitohaina[a], Cyril Grouin[a], Pierrick Bedouch[b,e], Georgeta Bordea[d], Amélie Daveluy[c], Vincent Depras[c], Natalia Grabar[a,f], Ghada Miremont-Salamé[c], Fleur Mougin[d], Cécile Pageot[c], Frantz Thiessard[c,d], Thierry Hamon[a,g]

[a] LIMSI, CNRS, Université Paris-Saclay, Campus universitaire d'Orsay, F-91405 Orsay cedex, France,
[b] CHU Grenoble Alpes & TIMC-IMAG UMR 5525 CNRS, Université Grenoble Alpes, CS 10217, F-38043 Grenoble cedex 9, France,
[c] CHU Bordeaux, Place Amelie Rabat Leon, F-33000 Bordeaux, France,
[d] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, 146 rue Léo Saignat, F-33000 Bordeaux, France
[e] CNHIM, 75-77 rue du Père Corentin, CS 91425, F-75993 Paris cedex 14, France,
[f] STL, CNRS, Université de Lille, Domaine du Pont-de-Bois, BP 60149, F-59653 Villeneuve-d'Ascq cedex, France,
[g] Université Paris 13, avenue Jean-Baptiste Clément, F-93430 Villetaneuse, France

Résumé

Dans cet article, nous présentons le schéma d'annotation utilisé pour étudier les interactions aliments-médicaments (Food-drug interaction - FDI). Le corpus se compose de 639 résumés d'articles scientifiques issus de Medline. Nous avons défini un schéma d'annotation constitué de 21 catégories d'entités et de 21 types de relations appliquées sur 9 catégories d'entités. Ces schémas ont été appliqués sur des documents rédigés en anglais ou en français, ouvrant la voie à un corpus multilingue annoté au moyen des mêmes catégories. Nous présentons également quelques expériences d'identification automatique des types de relations. L'adaptation de domaine à partir des interactions médicament-médicament (DDI) permet d'avoir un schéma d'annotation des relations selon 4 types. L'extraction automatique de ces relations conduit à une F1-mesure de 0.79 obtenue avec un modèle SVM précédé d'un processus de sélection de descripteurs SFM.

Mots-clés:

Interactions aliments-médicaments; Traitement Automatique des Langues; Classification.

Introduction

Les interactions entre médicaments et aliments *(Food-drug interaction - FDI)* constituent une information essentielle pour les médecins, afin d'assurer l'efficacité des traitements médicaux proposés aux patients. En effet, ces interactions peuvent réduire ou augmenter l'efficacité des traitements (nécessitant de revoir la posologie), produire des effets secondaires (notamment des effets secondaires inconnus), mais également modifier la manière dont le corps réagit face à une molécule (pharmacocinétique). Certains aliments sont connus pour les interactions qu'ils entretiennent avec les médicaments [1]. Ainsi, le pamplemousse augmente la fréquence et la gravité des effets secondaires dans les traitements contre le cholestérol ; les légumes verts à haute teneur en vitamine K (brocolis, chou vert, épinards, etc.) doivent être évités en parallèle de la prise d'anti-coagulants oraux ; et dans le cas des diurétiques, qui augmentent le taux de potassium dans le corps, il est recommandé d'éviter les aliments riches en potassium (bananes, oranges, etc.). L'accès à ces connaissances, notamment par le biais d'outils du traitement automatique des langues (TAL), est donc primordial pour adapter les traitements et mieux conseiller les patients.

Des campagnes d'évaluation ont porté sur les interactions entre médicaments, telles que la campagne Drug-Drug Interaction (SemEval 2013) sur l'anglais [2], donnant lieu à la production de corpus annotés partagés [3]. Concernant les interactions entre aliments et médicaments, des travaux ont notamment porté sur le pamplemousse [4] dont les interactions avec des médicaments sont désormais connues et documentées.

Dans cet article, nous présentons les principes d'annotations en entités et en relations que nous avons retenus pour annoter un corpus de résumés d'articles scientifiques du domaine médical. L'intérêt de cette démarche repose sur l'association d'informaticiens spécialisés en traitement automatique des langues (TAL), de pharmacologues d'un centre régional de pharmacovigilance (CRPV) et d'une association professionnelle (CNHIM) pour produire ces ressources. Nous décrivons également les expériences d'identification automatique des relations entre aliments et médicaments que nous avons menées sur la base de ce corpus annoté.

Corpus

Présentation

Le corpus se compose de 639 résumés d'articles scientifiques sur les interactions aliments-médicaments avec effet secondaire, issus de la base Medline. Ce corpus, nommé POMELO [4], a été produit à partir de la requête suivante dans PubMed : ("FOOD DRUG INTERACTIONS"[MH] OR "FOOD DRUG INTERACTIONS*") AND ("adverse effects*"). Chaque résumé comprend entre 1 et 659 mots, du simple mot-clé isolé au résumé structuré complet, répartis sur 2 à 10 lignes. Nous avons éliminé de ce corpus les documents ne contenant aucune interaction aliment-médicament, ceux comportant trop peu d'informations (uniquement le titre ou

des mots-clés, nombre de mots inférieur à dix), et ceux pour lesquels des informations clés sont absentes (période de prise d'un médicament par rapport aux repas, cinétique du médicament, etc.).

Processus d'annotation

Le schéma d'annotation s'inspire du guide d'annotation produit pour le corpus POMELO, que nous avons simplifié et complété. Dans un premier temps, nous avons travaillé sur un sous-ensemble de 20 fichiers composés de 79 à 406 mots. Afin d'accélérer le processus d'annotation et de gagner en qualité, une première phase d'annotation manuelle a été réalisée par les informaticiens en TAL, avec propagation automatique de ces annotations [5] (par exemple, la catégorie Drug associée au mot "Simvastatine" est automatiquement reportée sur toutes les autres occurrences de ce mot dans le corpus). Une double annotation indépendante a été réalisée par deux équipes de pharmacologues (CNHIM et CRPV) sur la base de cette pré-annotation pour corriger et compléter les annotations. Nous avons ensuite réalisé une phase de consensus entre les deux versions produites, en mobilisant les compétences des informaticiens spécialistes du TAL, ayant une vision de la représentation des données et des possibilités algorithmiques, et les pharmacologues, ayant les connaissances médicales. Sur cette base, nous avons finalisé le guide d'annotation en entités.

Le premier principe d'annotation retenu consiste à indiquer, pour chaque résumé, s'il est pertinent pour l'étude des interactions aliments-médicaments. Ce travail repose sur une analyse du contenu et permet d'écarter les documents dont les interactions n'impliquent aucun aliment (par exemple, lorsque la caféine est prise comme un médicament plutôt que consommée comme aliment). L'intérêt de l'annotation en pertinence est double : *(i)* seuls les documents jugés pertinents sont ensuite annotés en entités et relations, évitant de ce fait un travail d'annotation inutile, et *(ii)* cette annotation permet de s'assurer qu'aucun document n'a été oublié par les annotateurs.

Annotations des entités

Nous avons organisé notre guide d'annotation des entités autour de cinq grands domaines, selon que les entités relèvent : *(i)* du protocole expérimental, *(ii)* des pathologies, *(iii)* des aliments, *(iv)* des médicaments, ou *(v)* de l'anatomie. Un ensemble d'informations complémentaires est commun aux aliments et aux médicaments. Une dernière catégorie permet à l'annotateur d'indiquer des informations jugées utiles mais non prévues dans le guide.

Autour du protocole expérimental

Deux catégories renseignent de la manière dont l'étude scientifique a été conduite et de la population étudiée :

- "population" pour la population visée par l'étude *(adulte, enfants de moins de 6 ans, 15 hommes, rats mâles et femelles, personnes saines)*,
- "studyType" pour le type d'étude scientifique *(étude croisée double aveugle randomisée)*.

Autour des aliments

Quatre catégories concernent les aliments et leur prise :

- "food" pour les aliments de base ou transformés incluant le mode de préparation *(eau, orange, jus de fruits, thon, bœuf braisé ou grillé, thé infusé)*, y compris les suppléments alimentaires *(supplément de nutrition entérale)*,
- "component" sur les composants alimentaires qui ne sont pas pris comme compléments ou suppléments alimentaires *(acide nitrique, histamine, vitamine K)*,
- "mealTime" pour les informations temporelles de prise par rapport aux repas *(avant, à jeun, entre les repas)*,
- "meal" pour le type de régime alimentaire *(riche en graisses, petit déjeuner riche en calories, légumes riches en vitamine K, régime contrôlé)*.

Autour des pathologies

Six catégories concernant à la fois les problèmes pathologiques et les examens et résultats d'examen.

- "treatedDisease" pour la maladie traitée *(tuberculose)* ou l'opération chirurgicale envisagée *(opération de remplacement des valves)*,
- "pathology" pour les maladies, signes ou symptômes qui sont dûs à l'interaction aliment-médicament mais qui ne sont pas traités par un médicament *(diarrhée, perturbations non souhaitées, effet secondaire, pas de symptôme)*,
- "interactionMechanism" pour annoter les interactions et mécanismes observés dans le cadre de l'étude. Il s'agit essentiellement de noms, de verbes ou de participes passés *(augmentation, induction, inhibition, affecte, réduit, élevé)*. Un menu déroulant permet de déterminer la valence associée : négation, diminution, augmentation, excès,
- "parameter" pour annoter des éléments sur lesquels s'appliquent des concentrations ou des mesures *(concentration d'histamine, limite supérieure de l'intervalle thérapeutique, taux de clairance métabolique, valeur thrombo-test)*,
- "value" pour les valeurs numériques ou des qualificatifs, associés à un paramètre *(69%; 95% CI, 39-91%, de 0% à 42%)*,
- "exam" pour les examens avec une valeur associée *(scintigraphie de perfusion myocardique)*.

Autour des médicaments

Trois catégories concernent les médicaments :

- "drug" pour annoter les médicaments (dénomination commune internationale et noms commerciaux), y compris sous la forme d'abréviations *(simvastatine, tétracycline, Tcy, Zocor)*, leurs métabolites *(acide fusidique)* et les isomères,
- "drugEffect" pour annoter l'effet attendu *(anticoagulation, réponse hémodynamique)*. La classe pharmacologique des médicaments est également associée à cette catégorie puisque la classe renseigne de l'effet visé *(diurétique, traitement anticoagulant)*,
- "pharmacokinetics" pour annoter le devenir du médicament dans l'organisme ou les informations permettant de le mesurer *(absorption, clairance, activité du CYP2C9, oxydation du CYP3A, augmentation du C-MAX et de l'aire sous la courbe, pic de concentration, t1/2, AUC, C-MAX, INR)*.

Informations communes aux aliments et médicaments

Quatre catégories permettent d'apporter des informations complémentaires aux aliments et aux médicaments. Bien que ces catégories recouvrent le même type d'information, elles s'appliquent à des informations exprimées différemment selon qu'il s'agit d'aliment ou de médicament.

- "frequency" pour la fréquence de la prise d'aliment ou de médicament *(3 fois par jour, quotidiennement, consommation régulière)*,
- "dosage" pour la dose médicamenteuse *(200 ml, 0,4 mg/kg)* ou la quantité d'aliments *(apport élevé, un verre)*,
- "duration" pour la durée d'un traitement ou d'une consommation alimentaire *(pour 3 heures, période expérimentale de 7 jours)*,
- "mode" pour le mode d'administration *(oral, intra-veineux)* ou de consommation *(boisson, infusion, ingestion, ingéré, nourri)*.

Autour de l'anatomie

Deux catégories permettent l'annotation d'informations liées à l'anatomie, hors parties anatomiques qui n'ont pas été jugées pertinentes pour les interactions aliments-médicaments :

- "biologicalFunction" pour les fonctions biologiques *(insuffisance cardiaque, fraction d'éjection du ventricule gauche pauvre, pression sanguine)*,
- "enzyme" pour les enzymes *(CYP3A, cytochrome humain P450 (CYP) 2C9, glycoprotéine P, OATP)*.

Autres informations

Une dernière catégorie "other" permet à l'annotateur d'annoter des informations qu'il juge pertinentes mais qui ne correspondent à aucune des catégories existantes du guide. L'intérêt de cette catégorie réside dans la possibilité d'enrichir le guide d'annotation par les informations identifiées lors du travail d'annotation.

Application du schéma en corpus

Nous avons appliqué ce schéma d'annotation en entités sur des documents rédigés en anglais et en français. La figure 1 présente un extrait de corpus en français annoté en entités, dont les informations présentes permettent d'inférer que l'aliment *(ail)* réduit l'absorption du médicament *(saquinavir)*. Cet extrait correspond à un cas clinique décrit dans une thèse de pharmacie [6].

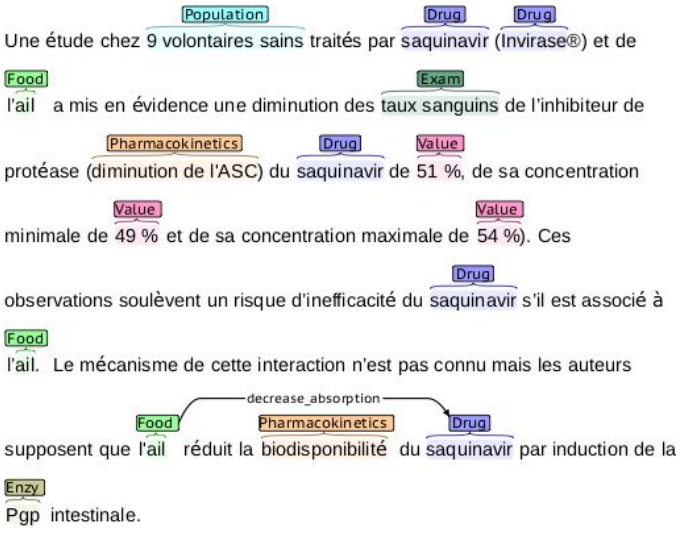

Figure 1 – Extrait de corpus annoté

Annotations des relations

Afin de mettre en place le schéma d'annotation des relations spécifiques aux interactions aliments-médicaments *(Food-drug interaction - FDI)*, et parce que le corpus de 639 résumés n'est pas encore disponible selon le schéma d'annotation en entités, nous avons défini les relations à partir d'une réflexion antérieure [4].

Nous avons ainsi annoté le corpus initial de 639 résumés selon 9 types d'entités (Drug, Food, MealTime, DrugEffect, TreatedDisease, SideEffect, Numbers, Frequency, Dosage, FoodSupplement) et 21 types de relations. Ce travail d'annotation a été réalisé avec Brat [7] par un externe en pharmacie. Les annotations se concentrent sur des informations sur la relation entre aliments, médicaments et pathologies. Étant donné que nous examinons les interactions aliment-médicament dans cet article, nous avons construit notre ensemble de données en tenant compte de tous les couples de drug et food ou food-supplement à partir des données POMELO. L'ensemble des données se compose de 902 phrases étiquetées parmi 13 types de relations autour des effets secondaires, de l'absorption et de l'élimination des médicaments :

- "Relation" (Rel) pour signaler une relation générique qu'il n'est pas possible de typer plus finement,
- "Improve drug effect" (Imp) si l'aliment améliore l'effet attendu du médicament,
- "Worsen drug effect" (Wors) si l'aliment aggrave l'effet attendu du médicament,
- "No effect on drug" (No) lorsque l'aliment ne produit aucun effet sur le médicament,
- "Positive effect on drug" (Pos) si l'aliment produit un effet positif sur le médicament,
- "Negative effect on drug" (Neg) si l'aliment produit un effet négatif sur le médicament,
- "New side effect" (New) si l'aliment produit un nouvel effet secondaire,
- "Increase absorption" (Inc) lorsque l'aliment augmente l'absorption du médicament,
- "Decrease absorption" (Dec) lorsque l'aliment diminue l'absorption du médicament,
- "Speed up absorption" (Spd), si l'aliment accélère l'absorption du médicament
- "Slow absorption" (Sl-a), si l'aliment ralentit l'absorption du médicament
- "Slow elimination" (Sl-e), si l'aliment ralentit l'élimination du médicament,
- "Without food" (Wout) si le médicament ne devrait pas être pris avec un aliment.

Les statistiques de l'ensemble de données sont fournies dans le tableau 1.

Tableau 1 – Distribution des annotations des relations

Catégorie de relation	Nombre	Pourcentage
Unspecified relation	530	58,8%
No effect on drug	109	12,1%
Decrease absorption	53	5,9%
Improve drug effect	6	0,7%
Positive effect on drug	21	4,8%
Negative effect on drug	88	9,8%
Speed up absorption	1	0,1%
Increase absorption	39	4,3%
Worsen drug effect	8	0,9%
Slow elimination	15	1,7%
Slow absorption	15	1,7%
New side effect	4	0,4%
Without food	13	1,4%
Total	902	100%

Adaptation au domaine à partir des interactions médicament-médicament (Drug-drug interaction - DDI)

Comme les interactions aliments-médicaments sont décrites de manière très fine selon de nombreux types de relations, nous sommes confrontés au manque de données et au manque d'exemples par type de relation comme indiqué dans le tableau 1. Par exemple, la relation "Speed up absorption" est représentée par un seul exemple, ce qui ne permet pas d'établir une méthode efficace pour extraire automatiquement la relation considérée. Par ailleurs, la tâche d'identification des FDIs présente des similitudes avec la tâche d'extraction des DDIs, où deux médicaments pris ensemble mènent à une modification de leurs effets. Dans cet article, nous étudions la projection des connaissances sur les DDIs afin d'obtenir de nouveaux types correspondant aux FDIs.

Correspondance de types DDI-FDI

Afin de vérifier la cohérence entre les deux domaines, nous proposons une approche permettant de trouver une correspondance de type DDI pour chaque type FDI. Pour ce faire, un modèle est entraîné sur les données DDI. Chaque type de relation FDI est ensuite représenté par un ensemble de descripteurs, formant ainsi une instance par type de relation. Puis, nous projetons le modèle entraîné avec les données DDI sur cette représentation afin de déterminer à quel type de DDIs correspond la relation. L'ensemble des données de relation est donc un vecteur $DR=[D_1, D_2, ..., D_n]$ de taille n, où n est le nombre de types de relations de type FDI, et D_i est un ensemble de descripteurs représentant la $i^{ème}$ relation. Dans nos expériences, nous construisons les descripteurs D_i à partir des descripteurs de chaque phrase P_i étiquetée par la relation R_i dans l'ensemble de données initial D. Ainsi, le modèle affecte une étiquette de type DDI pour chaque type FDI.

Contraste des étiquettes d'instances

Dans cette section, nous analysons individuellement les étiquettes DDI affectées aux instances FDI afin de déterminer si tous les membres d'une classe FDI donnée sont affectés au même type DDI et d'appuyer la correspondance obtenue. Pour ce faire, nous retenons le meilleur modèle obtenu pour la classification des DDIs et nous l'appliquons sur les données FDI afin d'obtenir de nouvelles étiquettes pour chaque instance. Cette méthode nous permet de contraster les annotations des données POMELO et une annotation basée sur les types DDIs. Une méthode de classification des instances FDI est ensuite mise en place en utilisant les étiquettes DDI.

Expériences

Dans cette section, nous présentons les expériences d'identification automatique des relations que nous avons réalisées au moyen de plusieurs classifieurs. Cette classification ne vise que les interactions aliments-médicaments.

Modèle DDI

L'adaptation de domaine DDI-FDI est évaluée à travers l'utilisation de classifieurs et de descripteurs.

Corpus – Le corpus DDI est composé de 2280 instances extraites de la base DrugBank et des résumés Medline, similaire au corpus utilisé lors de la compétition Drug-Drug Interaction 2013 [2], étiquetées en quatre types : *(i)* conseil (540 instances) pour une recommandation concernant l'utilisation concomitante de deux médicaments, *(ii)* effet (591) pour l'effet du DDI, *(iii)* mécanisme (1006) pour la pharmacodynamique (les effets d'un médicament sont modifiés par la présence d'un autre médicament) ou la pharmacocinétique (les processus par lesquels les médicaments sont absorbés, distribués, métabolisés et excrétés), et *(iv)* interaction (143) où aucune information sur l'interaction n'est fournie.

Descripteurs – Dans chaque phrase, les chiffres sont remplacés par le caractère "#" comme proposé par Kolchinsky [8], les autres caractères spéciaux sont supprimés, et chaque mot est converti en minuscules. Nous analysons l'impact de différents descripteurs : formes fléchies (F) des mots, lemmes (L), étiquettes morpho-syntaxiques (Po), fenêtres de mots précédant le premier argument de la relation (P), mots entre les deux arguments (E) et mots suivant le second argument (S).

Classification – La performance des classifieurs est évaluée selon la précision (P), le rappel (R), et la F-mesure (F1) obtenus par un processus de validation croisée en 10 échantillons. Nous utilisons l'implémentation Scikit-learn [9] des classifieurs : *(i)* un arbre de décision (DTree), *(ii)* un classifieur SVM *l2*-linéaire (LSVC-*l2*), *(iii)* une régression logistique (LogReg), *(iv)* un classifieur bayésien naïf multinomial (MNB), *(v)* une forêt aléatoire (RFC) et *(vi)* un SVM combiné avec un algorithme de sélection de descripteurs (SFM-SVM).

Adaptation de domaine

Afin d'étiqueter les données FDI selon les types DDI, la meilleure configuration obtenue est appliquée sur les données POMELO selon la méthode précédemment décrite. D'après les nouvelles étiquettes, on obtient une répartition différente des instances (voir tableau 4) : conseil (72 instances), effet (329), mécanisme (397), interaction (104). Les modèles de classification sont alors évalués sur nos données en utilisant les nouvelles étiquettes afin de déterminer l'effet de l'adaptation de domaine sur l'identification automatique des relations. De la même manière que pour le modèle DDI, nous évaluons les six classifieurs par un processus de validation croisée en 10 échantillons en faisant varier les descripteurs utilisés.

Résultats

Annotation manuelle des entités

Evaluation par rapport à la pré-annotation initiale

Le tableau 2 fournit les nombres de vrais positifs (VP), faux positifs (FP), faux négatifs (FN), ainsi que les valeurs de précision (P), rappel (R), et F-mesure (F) entre la version annotée par chaque équipe et la pré-annotation initiale.

Tableau 2 – Evaluation globale des annotations par rapport à la pré-annotation manuelle du corpus

Annotateur	VP	FP	FN	P	R	F1
CNHIM	577	9	40	0,9846	0,9352	0,9593
CRPV	515	71	419	0,8788	0,5514	0,6776

Distribution finale des annotations

Le tableau 3 fournit la distribution des annotations en entités sur le sous-corpus de 20 fichiers, après la recherche de consensus, classées par nombre d'annotations décroissant.

Tableau 3 – Distribution des annotations en entités

Catégorie d'entité	Nombre	Pourcentage
Drug	146	15,6%
Food	132	14,1%
Pharmacokinetics	117	12,5%
InteractionMechanism	100	10,7%
Value	45	4,8%
Dosage	45	4,8%
Component	44	4,7%
Parameter	42	4,5%
Mode	42	4,5%
MealTime	40	4,3%
Population	36	3,9%
Meal	30	3,2%
DrugEffect	21	2,2%
Enzyme	20	2,1%
StudyType	15	1,6%
Function	14	1,5%
Duration	14	1,5%
Frequency	13	1,4%
Pathology	10	1,1%
TreatedDisease	4	0,4%
Exam	2	0,2%
Other	2	0,2%

Annotation des relations

La figure 2 fournit les performances des différents classifieurs sur la tâche d'identification automatique des interactions médicament-médicament (DDI).

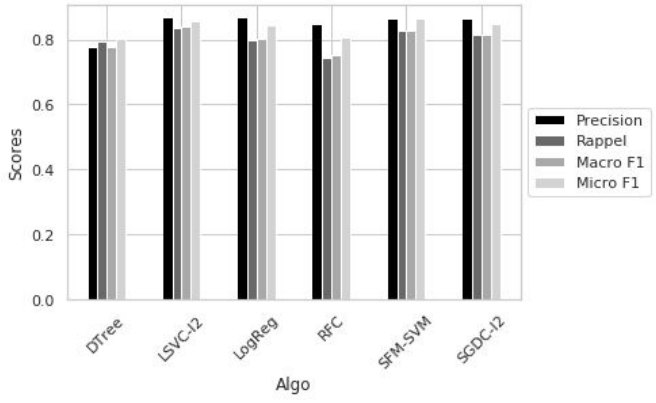

Figure 2 – Performances des modèles d'extraction automatique des relations obtenues sur les données DDI

Le tableau 4 présente les résultats de l'application du modèle DDI sur les données POMELO. La première ligne fournit la correspondance de types et les quatre dernières fournissent les pourcentages d'instances FDI affectées au type DDI, avec les types FDI : Relation (Rel), Decrease absorption (Dec), No effect on drug (No), Increase absorption (Inc), Negative effect on drug (Neg), Positive effect on drug (Pos), New side effect (New), Without food (Wout), Improve drug effect (Imp), Slow elimination (Sl-e), Slow absorption (Sl-a), Worsen drug effect (Wors), Speed up absorption (Spd) ; et les types FDI : Conseil (C), Mécanisme (M), Effet (E), Interaction (Int).

Tableau 4 – Correspondance type DDI-FDI (ligne 1) et pourcentage d'instances FDI affectées au type DDI (lignes 2–5)

FDI	Rel	Dec	No	Inc	Neg	Pos	New
DDI	M	M	M	M	E	E	E
Conseil	7	2	9	5	9	33	0
Effet	44	2	19	5	53	38	75
Int	17	2	1	0	16	0	0
Méca	32	94	71	90	22	29	25

FDI	Wout	Imp	Sl-e	Wors	Sl-a	Spd
DDI	C	E	M	E	M	M
Conseil	54	16	0	0	0	0
Effet	23	67	7	50	0	0
Int	0	0	0	0	0	0
Méca	23	17	93	50	100	100

Afin d'évaluer l'efficacité de l'adaptation de domaine sur nos données, nous avons appliqué les étiquettes DDI sur les données POMELO et effectué une identification automatique des relations selon un modèle de classification. La figure 3 présente les performances d'un modèle SVM avec une régularisation *l2* précédé d'un processus de sélection de descripteurs SFM, en utilisant les unigrammes, bigrammes et trigrammes des descripteurs : forme fléchie (F), lemme (L), lemmes précédents (P), entre (E) ou après (A) les arguments de la relation, et la catégorie morpho-syntaxique (Pos).

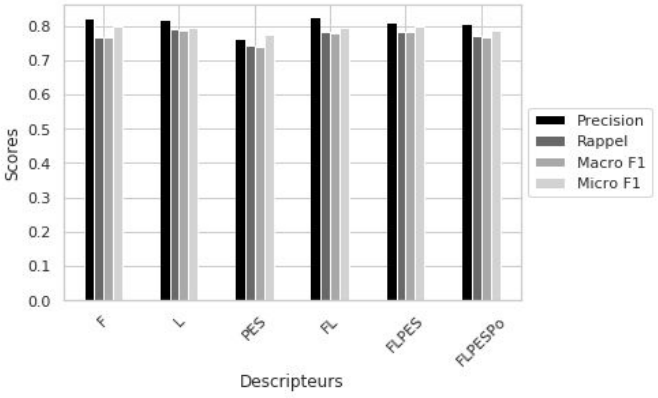

Figure 3 – Performance des modèles d'extraction automatique des relations sur les données POMELO en utilisant les étiquettes DDI selon les descripteurs

Discussion

Annotation manuelle des entités

Sur le sous-ensemble de 20 fichiers, deux fichiers ont été jugés non pertinents car ne comprenant aucune interaction

impliquant des aliments. Ils n'ont donc pas été annotés. Les dix-huit fichiers annotés résultant du consensus comprennent 934 entités. Quatre catégories sont majoritairement utilisées (drug, food, pharmacokinetics, interactionMechanism) et représentent plus de la moitié des annotations produites (53%).

D'après le tableau 2, qui présente l'évolution des annotations par rapport à la pré-annotation initiale, le faible nombre de faux positifs montre la qualité élevée de la pré-annotation, qui se traduit par des valeurs élevées de précision. Le nombre important de faux négatifs pour le CRPV (419) et la valeur moyenne de rappel pour cette équipe d'annotateurs (0,5514) témoigne d'une pré-annotation incomplète. Les ajouts effectués concernent principalement trois catégories d'entités : Pharmacokinetics (complexe et absente de la pré-annotation manuelle), Component et Pathology (incomplètes en raison d'une forte variété d'occurrences entre fichiers).

Nous avons également calculé l'accord inter-annotateurs entre les deux équipes d'annotateurs au moyen du coefficient Kappa. Nous rappelons que les deux équipes ont travaillé sur une base commune issue de la pré-annotation manuelle. En conséquence, une partie des accords observés est due à cette base commune. Nous calculons un coefficient Kappa global de 0,529, soit une absence d'accord. Dans le détail, et comme révélé par le tableau 2, il existe un écart important dans le nombre total d'annotations produites par chaque équipe (617 pour le CNHIM et 934 pour le CRPV) suffisant à expliquer cette absence d'accord. Une évaluation par la F-mesure révèle cependant un accord sur les deux principales catégories : Drug (F=0,8329) et Food (F=0,8765). Des désaccords plus importants sont constatés sur Pharmacokinetics (F=0,2917) et InteractionMechanism (F=0,5802), notamment en raison de la spécialité de chaque équipe, le CRPV ayant largement contribué aux annotations concernant la pharmacocinétique. D'autres explications plus habituelles lors de l'annotation de corpus sont également observées : oublis, erreurs de frontières, erreurs de catégories d'entités tant que le guide n'est pas finalisé.

Annotations des relations : Adaptation de domaine

Pour l'extraction des DDIs, la meilleure F1-mesure 0,839 est obtenue avec un SVM linéaire qui présente un bon rappel (figure 2) en utilisant les unigramme, bigramme et trigramme des lemmes comme descripteurs. Ce résultat est meilleur que celui obtenu par le meilleur modèle sur la même tâche lors de la compétition Semeval 2013 [2].

Nous avons appliqué ce modèle d'extraction des DDIs sur les représentations des relations FDI afin d'obtenir une correspondance entre les types DDI et FDI. Les résultats (tableau 4, ligne 1) indiquent une cohérence entre les deux domaines. En effet, les interactions impliquant la cinétique des médicaments (absorption, élimination, métabolisme) sont étiquetées "Mécanisme", les interactions impliquant les effets du médicament (positif, négatif, secondaire) sont étiquetées "Effet", la relation "Without food" qui est une contre-indication de prise d'aliment avec le médicament est étiquetée "Conseil". Cette cohérence des étiquettes démontrent l'efficacité de l'approche par la représentation des relations. Toutefois, nous remarquons que la relation non-spécifiée n'a pas été étiquetée "Int" (interaction). L'analyse des instances de cette relation (tableau 4, lignes 2-5) montre qu'un peu moins de la moitié est considérée comme des effets et le tiers comme des mécanismes. Ces informations pourront servir d'appui pour ajouter des précisions sur les interactions non spécifiées.

Nous remarquons que les instances impliquant des mécanismes sont étiquetées de manière relativement homogène. Ce phénomène peut être expliqué par la présence de mot-clés propres à ce type de relation comme "absorption" ou "elimination" tandis que le vocabulaire utilisé pour les relations d'effets ou de conseil sont beaucoup plus générique, ce qui provoque un effet d'ambiguïté sur le classifieur.

Avec les nouvelles étiquettes, le résultat obtenu sur la tâche d'extraction automatique des FDIs confirme effectivement l'efficacité de la méthode car la F-mesure est passée de 0,41 sur les étiquettes initiales à 0,79 sur les nouvelles étiquettes (figure 3). Ce résultat est obtenu par un modèle SVM avec une régularisation *l2* précédé d'un processus de sélection de descripteurs SFM, en utilisant les unigramme, bigramme et trigramme des lemmes comme descripteurs. La précision et le rappel sont visiblement équilibrés, ainsi que les macro et micro F-mesure (figure 3), ce qui suggère un bon équilibre des données.

Conclusion

Dans cet article, nous avons présenté les schémas d'annotations en entités et en relations que nous avons définis dans le cadre d'une étude sur les interactions aliments-médicaments *(Food-drug interaction - FDI)*. Ce travail a été réalisé sur un corpus de 639 résumés d'articles scientifiques parus en anglais. Nous avons également appliqué ces schémas sur des cas cliniques rédigés en français, sans avoir besoin d'adapter les schémas, ouvrant la voie à un corpus multilingue annoté au moyen des mêmes catégories. Pour l'annotation de relations, nous appliquons une adaptation de domaine à partir des interactions médicaments-médicaments (DDI) qui est une tâche similaire, afin d'établir une correspondance entre les types de relations et étiqueter les instances FDI selon les types DDI. Les résultats obtenus suggèrent que l'approche basée sur la représentation des relations à partir des instances est efficace pour identifier les correspondances des types et évaluer la cohérence entre les 2 domaines. Les nouvelles étiquettes obtenues peuvent servir d'appui pour préciser les interactions non spécifiées et pré-annoter de nouvelles données. L'adaptation de la méthode sur des corpus français suppose d'avoir recours à des données en français aussi bien sur les interactions DDI que FDI.

Les travaux futurs consistent à annoter l'ensemble du corpus anglais en entités, en s'aidant d'approches par apprentissage statistique, puis à annoter automatiquement les relations typiques des interactions aliments-médicaments sur la base des annotations en entités qui auront été réalisées. Nous envisageons également d'appliquer ces schémas d'annotation en entités et relations sur davantage de données en français. L'accès à ce type de ressources constitue cependant un frein important au développement de méthodes de TAL pour la langue biomédicale en français.

Remerciements

Ce travail a été financé par l'ANR dans le cadre du projet MIAM (référence ANR-16-CE23-0012).

Références

[1] Avoid Food-Drug Interactions. A Guide from the National Consumers League and U.S. Food and Drug Administration. https://www.fda.gov/media/79360/download

[2] I. Segura-Bedmar, P. Martínez, M. Herrero-Zazo, SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013).

[3] I. Segura-Bedmar, P. Martínez, M. Herrero-Zazo, The DDI Corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *J Biomed Inform* **46**:5 (2013), 914–920.

[4] T. Hamon, V. Tabanou, F. Mougin, N. Grabar, F. Thiessard, POMELO: Medline corpus with manually annotated food-drug interactions, *Proc of BioNLP* (2017).

[5] C. Grouin. Controlled propagation of concept annotations in textual corpora, *Proc of LREC* (2016).

[6] L. Aigueperce. Plantes à l'officine : soyons phytovigilants. Thèse de pharmacie, Faculté de Pharmacie de Grenoble (2014).

[7] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, J. Tsujii, BRAT: a Web-based Tool for NLP-Assisted Text Annotation, *Proc of Demonstration Session at EACL* (2012).

[8] A. Kolchinsky, A. Lourenço, H.Y. Wu, L. Li, L.M. Rocha, Extraction of pharmacokinetic evidence of drug–drug interactions from the literature. *PloS one* **10**:5 (2015), e0122199.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A . Passos, D. Cournapeau, M. Brucher, M. Perrot, E.Duchesnay, Scikit-learn, Machine-learning in Python. *Journal of Machine Learning Research* **12** (2011), 2825-2830.

Adresses de correspondance
Cyril Grouin, cyril.grouin@limsi.fr — Tsanta Randriatsitohaina, tsanta@limsi.fr