Identification des catégories de relations aliment-médicament

Tsanta Randriatsitohaina[1] Thierry Hamon[1,2]
(1) LIMSI, CNRS, Université Paris-Saclay, Campus universitaire d'Orsay, 91405 Orsay cedex, France
(2) Université Paris 13, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
tsanta@limsi.fr, hamon@limsi.fr

RÉSUMÉ

Les interactions aliment-médicament se produisent lorsque des aliments et des médicaments pris ensembles provoquent un effet inattendu. Leur reconnaissance automatique dans les textes peut être considérée comme une tâche d'extraction de relation à l'aide de méthodes de classification. Toutefois, étant donné que ces interactions sont décrites de manière très fine, nous sommes confrontés au manque de données et au manque d'exemples par type de relation. Pour résoudre ce problème, nous proposons une approche efficace pour regrouper des relations partageant une représentation similaire en groupes et réduire le manque d'exemples. Notre approche améliore les performances de la classification des FDI. Enfin, nous contrastons une méthode de regroupement intuitive basée sur la définition des types de relation et un apprentissage non supervisé basé sur les instances de chaque type de relation.

ABSTRACT

Identification of categories of food-drug relations.

Food-drug interactions occur when food and drug taken together cause unexpected effect. Their automatic identification from texts can be viewed as a relation extraction task relying on classification methods. Nevertheless, since these interactions are described in a very fine way, we are confronted with a lack of data and a lack of examples per type of relation. To solve this problem, we propose an efficient approach to cluster the relation sharing similar representation and to reduce the data sparseness. Our approach increases the efficiency of the FDI classification. Finally, when grouping the relations, we contrast an intuitive method based on the definition of the relation types, and an unsupervised learning relying on the examples associated with the FDI relation type.

MOTS-CLÉS : Classification, Clustering, Extraction de relation, Textes biomédicaux.

KEYWORDS: Classification, Clustering, Relation Extraction, Biomedical texts.

1 Introduction

Bien qu'il existe des bases ou des terminologies recensant les connaissances d'un domaine de spécialité, disposer d'informations à jour nécessite souvent d'avoir recours à l'analyse d'articles scientifiques. Ce constat est d'autant plus vrai lorsque les connaissances à recenser ne sont pas déjà présentes dans une base. Ainsi, si les interactions entre médicaments (Aagaard & Hansen, 2013) ou les effets indésirables d'un médicament (Aronson & Ferner, 2005) sont répertoriés dans des bases telles que DrugBank [1] ou Theriaque [2], d'autres informations comme les interactions entre un médicament

1. https://www.drugbank.ca/
2. http://www.theriaque.org

et un aliment y sont très peu présentes. Elles sont souvent fragmentées et dispersées dans des sources hétérogènes, principalement sous forme textuelle. Afin de répondre à ces problématiques de mises à jour ou de recensement de ces informations, des méthodes de fouille de textes sont généralement mises en œuvre (Cohen & Hunter, 2008; Rzhetsky *et al.*, 2009; Chowdhury *et al.*, 2011).

Dans cet article, nous nous intéressons à l'identification automatique de mentions d'interaction entre un médicament et un aliment (*Food-Drug Interaction* - FDI) dans des résumés d'articles scientifiques issus de la base Medline. A l'instar des interactions entre médicaments, une interaction entre un médicament et un aliment correspond à l'apparition d'un effet non attendu lors d'une prise combinée. Par exemple, le pamplemousse a un effet inhibiteur sur une enzyme impliquée dans le métabolisme de plusieurs médicaments (Hanley *et al.*, 2011). Pour extraire ces informations des résumés, nous faisons face à plusieurs difficultés : (1) les mentions des médicaments et des aliments sont très variables dans les résumés. Il peut s'agir des dénominations communes internationales ou des substances actives de médicaments tandis que pour les aliments, il peut s'agit d'un nutriment, d'un composant particulier ou d'une famille d'aliment ; (2) les interactions sont décrites de manière assez fine dans le corpus annoté à notre disposition, ce qui conduit à un faible nombre d'exemples. ; (3) nous disposons de résumés annotés avec des interactions aliment/médicament mais ces annotations ne couvrent pas de manière homogène tous les types d'interaction et l'ensemble d'apprentissage est bien souvent déséquilibré.

Nos contributions se concentrent sur l'extraction des FDI et l'amélioration des résultats des classifications en proposant une représentation des relations qui prend en compte le manque de données, en appliquant une méthode de regroupement sur le type de relations, et en utilisant des étiquettes de groupe dans une étape de classification afin d'identifier le type de FDI.

Après avoir présenté un état de l'art des méthodes d'acquisition de relations en corpus de spécialité (section 2), nous décrivons le corpus annoté que nous avons utilisé (section 3), notre approche pour regrouper les FDI annotées (section 4) et les expériences réalisées (5). Puis nous présentons et discutons les résultats obtenus (section 6) et nous concluons (section 7).

2 Etat de l'art

Différentes approches ont été proposées pour extraire les relations à partir des textes biomédicaux. Certaines approches combinent des patrons lexico-syntaxiques avec un CRF pour la reconnaissance des symptômes dans les textes biomédicaux (Holat *et al.*, 2016). D'autres approches génèrent automatiquement des données lexicales pour le traitement du texte libre dans les documents cliniques en s'appuyant sur un modèle alignement séquentiel multiple pour identifier des contextes similaires (Meng & Morioka, 2015). Toutefois, ces méthodes nécessitent des données bien fournies pour être efficace. Afin d'extraire les interactions médicament-médicament (DDI) (Kolchinsky *et al.*, 2015) se concentrent sur l'identification des phrases pertinentes et des résumés pour l'extraction des preuves pharmacocinétiques. (Ben Abacha *et al.*, 2015) proposent une approche basée sur du SVM combinant : (i) les caractéristiques décrivant les mots dans le contexte des relations à extraire, (ii) les noyaux composites utilisant des arbres de dépendance. L'union et l'intersection des résultats permettent d'obtenir des F1-mesures de 0,5 et 0,39 respectivement. Une méthode en deux étapes basée également sur le classifieur SVM est proposée par (Ben Abacha *et al.*, 2015) pour détecter les DDI potentiels, puis classer les relations parmis les DDI déjà identifiées. Cette seconde approche permet d'obtenir des F1-mesures de 0,53 et 0,40 sur des résumés Medline et 0,83 et 0,68 sur des documents de DrugBank. (Kim *et al.*, 2015) ont construit deux classifieurs pour l'extraction DDI : un classifieur binaire pour

extraire les paires de médicaments en interaction et un classifieur de types DDI pour identifier les catégories de l'interaction. (Cejuela *et al.*, 2018) considèrent l'extraction de la relation de localisation des protéines comme une classification binaire. (Liu *et al.*, 2016) proposent une méthode basée sur CNN pour l'extraction des DDI. Dans leur modèle, les mentions de médicaments dans une phrase sont normalisées de la manière suivante : les deux médicaments considérés sont remplacées par `drug1` et `drug2` respectivement dans l'ordre de leur apparition, et tous les autres médicaments sont remplacés par `drug0`. D'autres travaux utilisent un modèle de réseau neuronal récurrent avec plusieurs couches d'attention pour la classification DDI (Yi *et al.*, 2017; Zheng *et al.*, 2017), ou utilisant des récurrences au niveau des mots et des caractères (Kavuluru *et al.*, 2017) produisant une performance de 0,72. Sun *et al.* (2019) propose une méthode hybride combinant un réseau de neurone récurrent et convolutionnel induisant une amélioration de 3%. Le réseau convolutionnel profond de Dewi *et al.* (2017) permet de couvrir de longues phrases qui ont des jeux de données de DDI typique et obtenir une performance de 0,86.

Bien qu'efficace, ces méthodes sont difficilement applicable sur notre tâche car nos données souffrent d'un manque d'exemples par type de relation, ce qui nous a conduit à proposer une méthode de groupement pour y remédier.

3 Corpus

Des études permettant la constitution du corpus POMELO sur les interactions aliment-médicament ont déjà été menées (Hamon *et al.*, 2017). Ce corpus regroupe 639 résumés d'articles scientifiques du domaine médical (269 824 mots, 5 752 phrases), collectés à partir du portail PubMed [3] avec la requête : `("FOOD DRUG INTERACTIONS"[MH] OR "FOOD DRUG INTERACTIONS*" ) AND ("adverse effects*")`. Les 639 résumés sont annotés selon 9 types d'entités (Par exemple `DrugEffect`, `Treated disease`, `Side effect`, `Frequency`, `Dosage`) et 21 types de relation par un étudiant en pharmacie. Les annotations se concentrent sur des informations relatives aux aliments, médicaments et pathologies, ainsi qu'aux relations qu'ils entretiennent.

Puisque nous nous intéressons aux interactions aliment-médicament, notre ensemble de données est construit en retenant tous les couples de *drug* et *food* ou *food-supplement* à partir des données POMELO. L'ensemble de données qui en résulte est composé de 902 phrases étiquetées à l'aide de 13 types de relations : *decrease absorption* (5,9 %), *slow absorption* (1,7 %), *slow elimination* (1,7 %), *increase absorption* (4,3 %), *speed up absorption* (0,1 %), *new side effect* (0,4 %), *negative effect on drug* (9,8 %), *worsen drug effect* (0,9 %), *positive effect on drug* (2,3 %), *improve drug effect* (0,7 %), *no effect on drug* (12,1 %), *without food* (1,4 %), *unspecified relation* (58,8 %).

4 Groupements des types de relation

Comme indiqué dans la section 3, la distribution de notre ensemble de données est très déséquilibrée. Une description très détaillée des relations explique le manque d'exemples nombreux. Ainsi, la relation *speed up absorption* n'a qu'un seul exemple. Il n'est donc pas possible d'obtenir une bonne généralisation de la relation représentée. Pour résoudre ce problème, nous proposons deux méthodes

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

visant à regrouper les relations partageant des similarités afin d'obtenir plus d'exemples pour chaque groupe de relations. La première méthode repose sur la définition des types de relations (groupement intuitif) tandis que la seconde s'appuie sur un apprentissage non-supervisé.

4.1 Groupement intuitif

Nous proposons un premier regroupement intuitif des types de relations aliment-médicament. La tâche d'identification des FDI présente une similitude avec celle relative aux DDIs qui peuvent être regroupées selon l'absorption, la distribution, le métabolisme et l'excrétion (ADME) du médicament (Doogue & Polasek, 2013). Le regroupement intuitif des exemples est donc réalisé sur le même principe en prenant en compte la définition des relations :

1. **Relation non-précisée**. La relation *unspecified relation* identifie les FDI mentionnées sans autre précision. Les exemples associés représentant plus de la moitié des instances de relations annotées dans notre corpus, nous les considérons comme un seul groupe.
2. **Aucun effet**. La relation *No effect on drug* décrit les FDI exprimant explicitement que l'aliment considéré n'a aucun effet sur le médicament. Ces instances sont donc également représentées comme un seul groupe.
3. **Réduction**. Les relations *decrease absorption*, *slow absorption*, *slow absorption*, *slow elimination* décrivent une diminution de l'action du médicament sous l'influence d'un aliment, ils sont regroupés sous la relation *reduction*.
4. **Augmentation**. De manière similaire à la relatoin précédente, les relations *increase absorption*, *speed up absorption* sont regroupées sous la relation *augmentation*.
5. **Négatif**. Le groupe de relations *negative* comprend les relations *new side effect*, *negative effect on drug*, *worsen drug effect*, *with without food*.
6. **Positif**. Par analogie avec la relation précédente, les relations *positive effect on drug*, *improve drug effect* sont regroupées sous la relation *positive*.

Nous notons cette méthode de regroupement intuitive, ARNP (Augmentation, Réduction, Négatif et Positif). Ainsi, nous obtenons 6 types de relations aliments-médicament et ainsi un nombre plus important d'exemples par type (voir tableau 1).

Type	#	Pourcentage	Type	#	Pourcentage
unspecified relation	476	57,3 %	*no effect on drug*	109	13,1 %
reduction	79	9,5 %	*augmentation*	39	4,7 %
negative	103	12,4 %	*positive*	25	3 %
Total			831		100 %

TABLE 1 – Nombre et pourcentage de relations obtenues par le regroupement intuitif

4.2 Apprentissage non-supervisée

Dans cette section, nous proposons d'utiliser une méthode d'apprentissage non-supervisé pour regrouper les relations impliquant un effet d'un aliment sur un médicament.

4.2.1 Représentation des relations avec sélection de descripteurs

Pour regrouper les relations aliment-médicament, nous les représentons à l'aide d'un ensemble de descripteurs D =[F_1, F_2, ..., F_n], où n est le nombre de relations à regrouper, F_i est un ensemble de descripteurs représentant la relation R_i. La façon la plus naturelle d'obtenir des descripteurs F_i est de regrouper toutes les phrases S_i étiquetées par relation R_i dans l'ensemble de données initial D_S : F_i = Concaténation(S_i) pour tout S_i dans D_S. Nous considérons cette représentation de base (*baseline*) de notre tâche. Cependant, afin d'améliorer la représentation des relations, nous proposons une représentation supervisée pour extraire les descripteurs les plus pertinents pour la relation R_i en entraînant un classifieur SVM n-classes sur l'ensemble de données initial D_S. Puisque la décision du SVM est basée sur un hyperplan qui maximise la marge entre les échantillons et l'hyperplan séparateur représenté par $h(x) = w^T x + w_0$, où $x = (\mathbf{x}_1, ..., \mathbf{x}_N)^T x = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ est le vecteur de descripteurs et $w = (\mathbf{w}_1, ..., \mathbf{w}_N)^T$ le vecteur des poids, nous pouvons déterminer, à partir de ces pondérations, l'importance de chaque descripteur grâce à la matrice de coefficients associés aux descripteurs C de taille $n \times nf$ où nf est le nombre de descripteurs.

Les nm descripteurs les plus importants sont extraits pour chaque relation afin de représenter la relation R_i à l'aide d'un vecteur de descripteurs qui correspond aux premiers descripteurs positives du i^{me} vecteur de C. L'ensemble de données résultant est une matrice D =[F_1, F_2, ..., F_n] de taille $n \times nm$, où F_i est le descripteur extrait pour représenter la relation i, n est le nombre de classes (ici relations) et nm est le nombre de descripteurs à extraire. Afin de mieux saisir l'expression de la relation dans une phrase, nous proposons d'utiliser comme descripteurs, les lemmes avant le premier argument de la relation, les lemmes entre les deux arguments, et les lemmes après le deuxième argument de la classification SVM. Nous notons cette méthode, BBA-SVM.

4.2.2 Groupement des relations et identification des catégories

Etant donnée la définition des relations (section 4.1), nous cherchons à regrouper automatiquement dans 4 groupes, les types de relation, à l'exception des types *unspecified relation* et *no effect on drug*. Nous appliquons l'approche proposée à la section 4.2.1 sur les phrases du corpus POMELO, étiquetées par les 11 types de relations précises. Les données résultantes sont une matrice D de taille 11 $\times nm$ où nm est le nombre de descripteurs représentant un type de relation. L'algorithme d'apprentissage non-supervisé les regroupe alors dans 4 clusters. Le résultat est un vecteur des étiquettes des clusters Cl =[Cl_1, Cl_2, ..., Cl_{11}] où Cl_i est le cluster auquel la relation R_i appartient. Une fois les clusters définis, les étiquettes des phrases de l'ensemble de données initial sont remplacées par les étiquettes des clusters associés à la relation, et une classification supervisée est appliquée afin d'identifier le type de FDI grâce aux étiquettes des regroupements (cf. figure 1).

5 Expériences

Pour déterminer le type de FDI, nos expériences sont axées sur la qualité de la classification des relations sur le corpus POMELO.

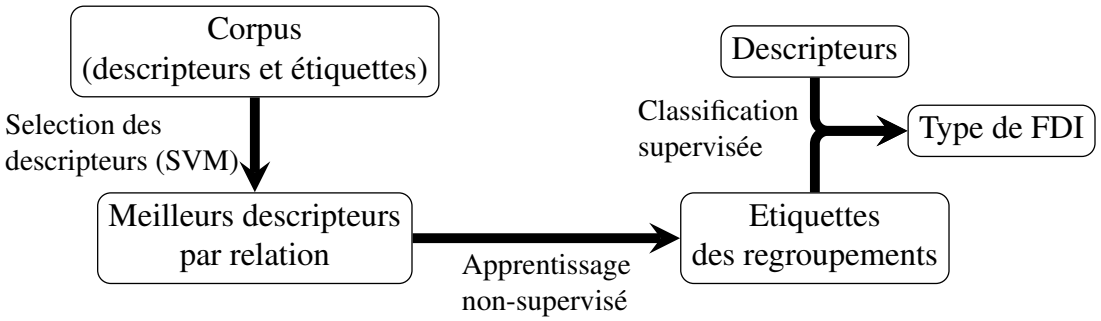

FIGURE 1 – Architecture de l'approche

5.1 Apprentissage non-supervisé

Représentation des relations L'impact de la représentation des relations sur les performances de la classification est évalué en faisant varier les descripteurs associés aux exemples : (1) l'ensemble de mots des phrases contenant la relation R (*baseline*) ; (2) les lemmes sont fournis à un classifieur SVM et les descripteurs les plus pertinents sont retenus (méthode lemme/SVM) ; (3) les lemmes précédant le premier argument de la relation, les lemmes entre les deux arguments, et les lemmes suivant le deuxième argument sont fournis à un classifieur SVM ; les meilleurs descripteurs sont sélectionnés (méthode BBA-SVM) ; (4) les descripteurs précédents auxquels s'ajoutent les formes fléchies des mots sont fournis à un classifieur SVM et les descripteurs les plus pertinents sont sélectionnés (ILBBA).

Algorithmes de groupement Pour évaluer notre approche, nous comparons les performances de 4 algorithmes d'apprentissage non-supervisé : (1) KMeans – les données sont divisées en k sous-ensembles, les k points centraux (centroïdes de partitions) sont identifiés de telle sorte que la distance entre le centroïde et les points à l'intérieur de chaque partition est minimale ; (2) Mini Batch K-Means – cette variante du KMeans utilise des sous-ensembles des données d'entrée, tirés au hasard à chaque itération de l'apprentissage ; (3) Un partitionnement spectral qui effectue une projection de faible dimension de la matrice de similarité des échantillons, suivie d'un KMeans ; (4) Une classification ascendante hiérarchique permettant de fusionner successivement les groupes d'instances.

Évaluation des groupements Nous utilisons 4 métriques pour évaluer l'assignation de la classification par rapport au groupement intuitif ARNP : (1) l'indice Rand ajusté qui mesure la similarité des deux assignations, sans tenir compte des permutations et avec normalisation aléatoire ; (2) l'homogénéité qui mesure la distribution des classes dans chaque groupe ; (3) la complétude qui mesure la distribution des groupes dans les classes ; (4) l'indice de Calinski-Harabaz pour évaluer le modèle, tel que plus un score Calinski-Harabaz est élevé, meilleure est la définition du modèle.

5.2 Classification des catégories FDI

Prétraitement. Chaque phrase est prétraitée comme suit : les chiffres ont été remplacés par le caractère'#' tel que proposé dans (Kolchinsky *et al.*, 2015), les autres caractères spéciaux sont supprimés, chaque mot est converti en minuscules.

5.2.1 Descripteurs

Pour évaluer l'efficacité de l'approche proposée, les descripteurs sont composés des formes fléchies, des lemmes, des catégories morpho-syntaxiques des mots, des lemmes précédant le premier argument de la relation, des lemmes entre les deux arguments et des lemmes suivant le second argument.

5.2.2 Modèles de classification

Dans nos expériences, nous évaluons la performance des classifieurs suivant issus de Scikit-learn : (1) un arbre de décision (DTree), (2) un classifieur SVM l2 linéaire (LSVC-l2), (3) une régression logistique (LogReg), (4) un classifieur bayesien naïf multinomiale (MNB), (5) un classifieur à base de forêt d'arbre de décision (RFC), et (6) un SVM combiné à une sélection des descripteurs (SFM-SVM).

5.2.3 Qualité de la classification

Puisque notre but est d'extraire les interaction aliment-médicament, nous devons évaluer notre approche par sa capacité à identifier de telles relations, celle-ci étant caractérisée par les scores d'évaluation du classifieur dans chaque expérience. Nous avons utilisé 3 mesures d'évaluation : précision (P), rappel (R), Score F1 (F_1). Considérant que l'un des enjeux de la tâche est la réduction du déséquilibre du nombre d'exemples par classe, nous considérons les macro-mesures d'évaluation qui calcule les mesures par classe, puis effectue la moyenne globale, et les micro-mesures, qui calculent les mesures sur l'ensemble des résultats indépendamment des classes. L'évaluation s'appuie sur un processus de validation croisée en 10 plis.

6 Résultats et discussion

Dans cette section, nous présentons les résultats obtenus avec les différentes configurations pour identifier les types de FDI à partir du corpus POMELO mais aussi les étiquettes de groupements obtenus automatiquement (tableau 2). Le meilleur résultat est obtenu lorsque 200 descripteurs sont sélectionnés par la méthode BBA-SVM, avec un partitionnement spectral pour le regroupement des types de relations, et un classifieur SFM-SVM utilisant comme descripteurs, les lemmes précédent le premier argument de la relation, les lemmes entre les deux arguments, et les lemmes suivant le second argument de la relation.

Représentation des relations	KMeans	MBKM	Spectral	Agglomerative
Baseline	0,362	0,394	**0,522**	0,374
Lemme	0,361	**0,405**	0,373	0,366
BBA-SVM	0,385	0,384	**0,58***	0,361
Formes fléchies+Lemme+BBA	0,473	0,507	0,367	**0,517**

TABLE 2 – Macro F-mesures obtenues à l'aide de différentes méthodes de représentation des relations et différents algorithmes d'apprentissage non-supervisé (*KMeans*, *MiniBatch-KMeans (MBKM)* , *Partitionnement spectral (Spectral)*, *classification ascendante hiérarchique (Agglomerative).*

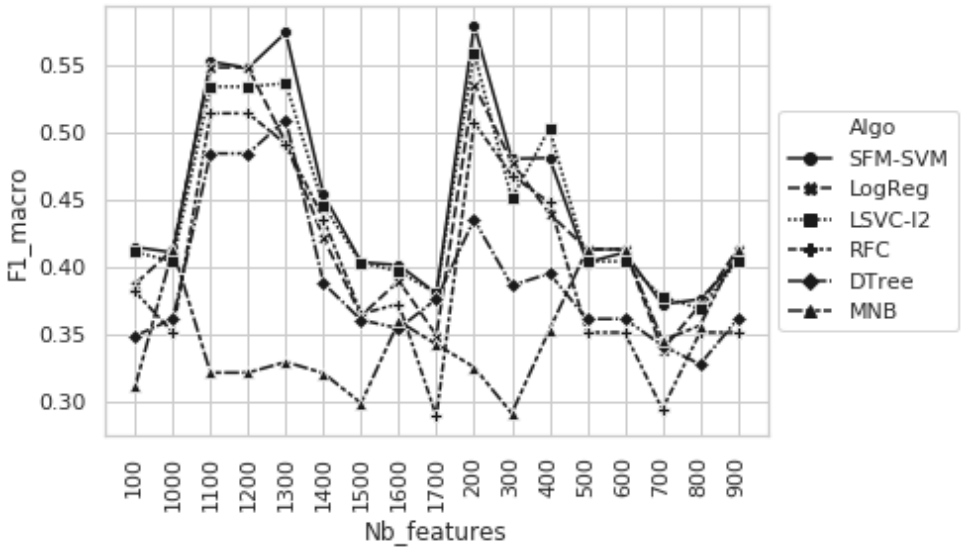

FIGURE 2 – Macro F-mesures obtenues sur les différents modèles en faisant varier le nombre de descripteurs utilisés pour représenter les relations lors de la classification des relations basée sur le meilleur modèle de regroupement des types de relations.

Descripteurs	Aucun regroupement	Méthode de regroupement ARNP	KMeans	MBKM	SpecC	AggloC
formes fléchies	0,379	0,378	0,416	**0,425**	0,573	0,422
formes fléchies et étiquettes morpho-syntaxiques	0,388	0,368	0,397	0,413	0,559	0,406
lemme	**0,41**	0,34	**0,462**	0,396	0,563	**0,449**
formes fléchies et lemmes	0,38	**0,401**	0,423	**0,425**	0,575	0,439
lemmes et étiquettes morpho-syntaxiques	0,403	0,392	0,444	0,386	0,56	0,436
formes fléchies et lemmes et étiquettes morpho-syntaxiques	0,387	0,395	0,425	**0,425**	0,567	0,443
BBA	0,364	0,364	0,385	0,384	**0,58***	0,361
formes fléchies et lemmes et BBA	0,383	0,384	0,394	0,402	0,576	0,403
formes fléchies et lemmes et BBA et étiquettes morpho-syntaxiques	0,379	0,396	0,383	0,398	0,566	0,404

TABLE 3 – Macro F-mesures obtenues avec différents types de decripteurs utilisés pour l'identification des relations selon la méthode de regroupement des types de relations utilisée.

		ARNP	Baseline	Lemme	BBA-SVM	ILBBA
Répartition des types de relations par regroupement	decrease absorption	C1	C1	C1	C1	C1
	improve drug effect	C2	C1	C3	C1	C3
	increase absorption	C3	C3	C1	C1	C1
	negative effect on drug	C4	C1	C1	C1	C1
	new side effect	C4	C1	C1	C4	C4
	positive effect on drug	C2	C4	C1	C1	C1
	slow absorption	C1	C1	C1	C1	C1
	slow elimination	C1	C2	C2	C1	C1
	speed up absorption	C3	C1	C2	C3	C1
	without food	C4	C1	C4	C1	C2
	worsen drug effect	C4	C1	C3	C2	C1
Évaluation des regroupements	Homogénéité	1,0	0,343	0,272	0,284	0,284
	Complétude	1,0	0,519	0,312	0,431	0,431
	Indice Rand ajusté	1,0	0,101	-0,132	-0,043	-0,043
	Indice de Calinski-Harabaz	1,0	0,376	0,999	0,903	1,146
Qualité de la classification	Macro précision	0,385	0,547	0,38	**0,589**	0,372
	Macro rappel	0,368	0,518	0,381	**0,582**	0,371
	Macro F-mesure	0,364	0,522	0,373	**0,58**	0,367
	Micro F-mesure	0,599	0,656	0,647	**0,67**	0,652

TABLE 4 – Répartition et évaluation des types de relations par regroupement obtenus à l'aide de différentes représentation des relations.

L'approche basée sur la représentation des relations BBA-SVM permet d'obtenir les meilleurs résultats avec une macro-F mesure de 0,58 pour identifier les FDI (tableau 2). On observe une différence de 0,23 entre les résultats obtenus grâce à un groupement ARNP et des données non-groupées (table 3). Ce résultat est obtenu en n'utilisant que 200 descripteurs pour regrouper les types de relations (figure 2), et à partir d'un classifieur SVM utilisant 1676 descripteurs composés des lemmes précédant le premier argument de la relation, des lemmes entre les deux arguments, et les lemmes après le deuxième argument. Ces résultats justifient notre hypothèse selon laquelle une relation se caractérise par des descripteurs spécifiques dans le contexte des arguments des relations. Ainsi, le fait que la méthode de sélection des descripteurs utilisé avant un classifier SVM (SFM-SVM) (table 5) permet d'obtenir de meilleures performances suggère que certains descripteurs sont plus importants que d'autres. Les sélectionner au préalable permet d'améliorer l'efficacité de la la prise de décision du classifieur. Nous observons également une différence entre les micro et macro F-mesure, qui passe de 0,23 avec l'ARNP à 0,09 (table 4). Celle-ci suggère une réduction du déséquilibre des données. De plus, la régression logistique permet d'obtenir les meilleurs résultats sur le micro F-mesure, mais celle-ci est un peu moins efficace au niveau de la macro F-mesure, ce qui signifie que le modèle est plus sensible au déséquilibre des données que les modèles basés sur un classifieur SVM.

En outre, le score élevé de Calinski-Harabaz (tableau 4) indique les regroupements sont plus denses et bien disjoints et montre l'efficacité de notre approche. Néanmoins, les autres mesures de regroupement indiquent une assignation indépendante par la méthode ARNP. L'analyse des étiquettes attribuées à chaque relation fournit une explication. Ainsi, dans le tableau 4, on observe aussi que trois relations (*worsen drug effect*, *speed up absorption* , *new side effect*) sont représentées individuellement tandis que toutes les autres sont regroupées dans un même cluster, sans qu'il n'y ait d'interprétation particulière expliquant l'existence de ce regroupement. Cela suggère que les 3 relations sont explicitement différentes des autres mais aussi que les autres relations ne sont pas suffisamment séparables. Il est également possible que le corpus POMELO contienne des annotations erronées. Ainsi, l'annotation pourrait être améliorée grâce à notre méthode de regroupement. Cette méthode de regroupement

Modèle	Précision	Rappel	macro F-mesure	micro F-mesure
DTree	0,45	0,441	0,435	0,614
LSVC-l2	0,572	0,558	0,56	0,668
LogReg	0,564	0,528	0,536	**0,674**
MNB	0,342	0,344	0,325	0,535
RFC	0,586	0,496	0,508	0,665
SFM-SVM	**0,589**	**0,582**	**0,58**	0,67

TABLE 5 – Mesures d'évaluation de l'identification des types des relations, obtenues en utilisant les descripteurs précédant, entre et suivant les arguments des relations et un algorithment de partitionnement spectral.

pourrait également être utile pour la classification des relations sans plus la précision (*unspecified relation*). En effet, ces données représentent plus de la moitié des relations, créant ainsi des ambiguïtés qui rendent l'identification des relations FDI difficile. Cependant, ces résultats montrent que notre approche conduit déjà à améliorer l'identification des types de FDI.

7 Conclusion et travaux futurs

Dans cet article, nous nous intéressons à l'identification des interactions aliment-médicament (FDI) dans la littérature scientifique. Lors de la mise en œuvre de méthodes d'apprentissage supervisé visant à identifier ces relations, nous faisons face au manque d'exemples dû au nombre élevé de types de relations. Pour résoudre ce problème, nous proposons de représenter chaque relation par les descripteurs les plus importants extraites à l'aide d'un classifieur SVM, puis nous regroupons les types de relations à l'aide d'un apprentissage non-supervisé. Les étiquettes des groupes sont ensuite utilisées comme étiquettes de relations sur l'ensemble de données initiales. Nous avons évalué l'utilité de ces regroupements en identifiant les FDI du corpus POMELO à l'aide d'algorithme d'apprentissage supervisé. Notre approche atteint les meilleures performances avec 200 descripteurs regroupées grâce à un partitionnement spectral, puis une extraction des relations FDI à l'aide d'un classifieur SVM. Nous obtenons une amélioration de la F-mesure de 0,23 par rapport à un regroupement intuitif et des relations non regroupées. En outre, la diminution de la différence entre les macro et micro f-mesures suggère une réduction du déséquilibre des données. Ainsi, les résultats des expériences confirment l'efficacité de notre approche. Pour les travaux futurs, nous envisageons l'identification du type de FDI basée sur une classification multi-étiquette, en utilisant les regroupements comme première étiquette. Nous souhaitons également tirer partie de classes ADME (Doogue & Polasek, 2013) (Absorption - Distribution - Métabolisme - Excrétion) et des modèles d'interaction médicament-médicament pour évaluer l'impact de ces informations à travers un apprentissage par transfert.

Remerciements

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

Références

AAGAARD L. & HANSEN E. (2013). Adverse drug reactions reported by consumers for nervous system medications in europe 2007 to 2011. *BMC Pharmacology & Toxicology*, **14**, 30.

ARONSON J. & FERNER R. (2005). Clarification of terminology in drug safety. *Drug Safety*, **28**(10), 851–70.

BEN ABACHA A., CHOWDHURY M. F. M., KARANASIOU A., MRABET Y., LAVELLI A. & 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.

CEJUELA J. M., VINCHURKAR S., GOLDBERG T., PRABHU SHANKAR M. S., BAGHUDANA A., BOJCHEVSKI A., UHLIG C., OFNER A., RAHARJA-LIU P., JENSEN L. J. & ROST B. (2018). LocText : relation extraction of protein localizations to assist database curation. *BMC Bioinformatics*, **19**(1), 15.

CHOWDHURY F. M., LAVELLI A. & MOSCHITTI A. (2011). A study on dependency tree kernels for automatic extraction of protein-protein interaction. In *Proceedings of BioNLP 2011 Workshop*, p. 124–133 : Association for Computational Linguistics.

COHEN K. & HUNTER L. (2008). Getting started in text mining. *PLoS Computational Biology*, **4**(1), e20.

DEWI I. N., DONG S. & HU J. (2017). Drug-drug interaction relation extraction with deep convolutional neural networks. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 1795–1802.

DOOGUE M. & POLASEK T. (2013). The abcd of clinical pharmacokinetics. *Therapeutic Advances in Drug Safety*, **4**, 5–7.

HAMON T., TABANOU V., MOUGIN F., GRABAR N. & THIESSARD F. (2017). Pomelo : Medline corpus with manually annotated food-drug interactions. In *Proceedings of Biomedical NLP Workshop associated with RANLP 2017*, p. 73–80, Varna, Bulgaria.

HANLEY M., CANCALON P., WIDMER W. & GREENBLATT D. (2011). The effect of grapefruit juice on drug disposition. *Expert Opin Drug Metab Toxicol*, **7**(3), 267–286.

HOLAT P., TOMEH N., CHARNOIS T., BATTISTELLI D., JAULENT M.-C. & MÉTIVIER J.-P. (2016). Weakly-supervised symptom recognition for rare diseases in biomedical text. In *Proceedings of the 15th International Symposium IDA 2016*, Lecture Notes in Computer Science, 9897, Advances in Intelligent Data Analysis XV, p. 192–203, Stockholm, Sweden.

KAVULURU R., RIOS A. & TRAN T. (2017). Extracting drug-drug interactions with word and character-level recurrent neural networks. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, p. 5–12 : IEEE.

KIM S., LIU H., YEGANOVA L. & WILBUR W. J. (2015). Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of biomedical informatics*, **55**, 23–30.

KOLCHINSKY A., LOURENÇO A., WU H.-Y., LI L. & ROCHA L. M. (2015). Extraction of pharmacokinetic evidence of drug–drug interactions from the literature. *PloS one*, **10**(5), e0122199.

LIU S., TANG B., CHEN Q. & WANG X. (2016). Drug-drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, **2016**.

MENG F. & MORIOKA C. (2015). Automating the generation of lexical patterns for processing free text in clinical documents. *Journal of the American Medical Informatics Association*, **22**(5), 980–986.

RZHETSKY A., SERINGHAUS M. & GERSTEIN M. B. (2009). Getting started in text mining : Part two. *PLoS Comput Biol*, **5**(7), e1000411.

SUN X., DONG K., MA L., SUTCLIFFE R. F. E., HE F., CHEN S.-S. & FENG J. (2019). Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy*, **21**, 37.

YI Z., LI S., YU J., TAN Y., WU Q., YUAN H. & WANG T. (2017). Drug-drug interaction extraction via recurrent neural network with multiple attention layers. In *International Conference on Advanced Data Mining and Applications*, p. 554–566 : Springer.

ZHENG W., LIN H., LUO L., ZHAO Z., LI Z., ZHANG Y., YANG Z. & WANG J. (2017). An attention-based effective neural model for drug-drug interactions extraction. In *BMC Bioinformatics*.