Extraction d'interactions entre aliment et médicament : Etat de l'art et premiers résultats

Tsanta Randriatsitohaina[1]
(1) LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay
tsanta.randriatsitohaina@limsi.fr

RÉSUMÉ

Dans cet article, nous nous intéressons à l'extraction des interactions entre médicaments et aliments, une tâche qui s'apparente à l'extraction de relations entre termes dans les textes de spécialité. De nombreuses approches ont été proposées pour extraire des relations à partir de textes : des patrons lexico-syntaxiques, de la classification supervisée, et plus récemment de l'apprentissage profond. A partir de cet état de l'art, nous présentons une méthode basée sur un apprentissage supervisé et les résultats d'une première série d'expériences. Malgré le déséquilibre des classes, les résultats sont encourageants. Nous avons ainsi pu identifier les classifieurs les plus performants suivant les étapes. Nous avons également observé l'impact important des catégories sémantiques des termes comme descripteurs.

ABSTRACT

Extraction of food-drug interactions :
State of the art and first results.

In this paper, we are interested in the extraction of food-drug interactions (FDI), a task which is similar to the extraction of relation between terms in specialized texts. Many approaches have been proposed for extracting relations from texts, including lexico-syntactic patterns, statistical learning-based methods, and more recently deep learning-based methods. From this state of the art, we present a supervised classification method and the results of a first set of experiments. Despite the imbalance of classes, the results are encouraging. We have identified the most relevant classifiers according to the steps of our methods. We have also observed the important impact of the semantic tags of terms used as features.

MOTS-CLÉS : Interaction aliment-médicament, Relation sémantique, Corpus de spécialité, Classification supervisée.

KEYWORDS: Food-Drug Interaction, Semantic relation, Specialized corpora, Supervised classification.

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 le recours à la consultation 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 et un aliment y sont très peu présentes et 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 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 quand leur prise est combinée. Par exemple, le pamplemousse est connu pour avoir un effet inhibiteur sur un enzyme impliqué dans le métabolisme de plusieurs médicaments (Hanley *et al.*, 2011). D'autres aliments peuvent avoir des effets sur l'absorption d'un médicament ou sur sa distribution dans l'organisme (Doogue & Polasek, 2013). Pour extraire ces informations des résumés, nous faisons face à plusieurs difficultés : (1) les médicaments et les aliments sont mentionnés de manière très variable dans les résumés. Il peut s'agir des dénominations communes internationales ou des substances actives de médicaments tandis que pour les aliments, il peut aussi être fait mention 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 nombre d'exemples peu conséquent. ; (3) bien que nous disposons de résumés annotés avec des interactions aliment/médicament, l'ensemble des annotations ne couvrent pas de manière homogène les différents types d'interaction et l'ensemble d'apprentissage est bien souvent déséquilibré.

Nous considérons l'extraction de ces interactions comme une tâche d'acquisition de relation étant donné les mentions d'un aliment et d'un médicament. Afin de répondre à ces difficultés, nous proposons d'utiliser une méthode de classification supervisée en quatre étapes : (1) sélection des phrases des résumés contenant les relations pertinentes ; (2) identification des grands types de relations correspondant aux annotations dans les phrases sélectionnées ; (3) la catégorisation des relations correspondant à des interactions entre aliment et médicament ; (4) identification des entités (aliment et médicament) en interaction.

Après avoir présenté un état de l'art des méthodes d'acquisition de relations en corpus de spécialité (section 2), nous décrivons à la section 3, le corpus annoté que nous avons utilisé pour mettre au point notre approche afin d'extraire des interactions entre médicament et aliment (section 4). Puis nous présentons et discutons les résultats obtenus lors d'expériences préliminaires (section 5) et nous concluons (section 6).

2 Etat de l'art

Depuis de nombreuses années, différents types d'approche ont été explorés pour extraire des relations à partir de textes.

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

Patrons lexico-syntaxiques Hearst (1992) a proposé une méthode à base de patrons lexico-syntaxiques pour l'extraction des hyponymes. Il s'agit de reconnaître les contextes discriminants pour les éléments en question et d'identifier les points communs entre plusieurs instances de la relation pour en faire émerger du corpus, les séquences d'éléments textuels ou de catégories morpho-syntaxiques caractéristiques d'une relation. Alors que cette première approche s'appuie sur des observations en corpus de langue générale pour identifier les patrons de la relation d'hyperonymie, il est également possible de la mettre en œuvre sur des textes de spécialité. Morin (1998) utilise les relations issues d'une terminologie comme exemple pour construire automatiquement des patrons lexico-syntaxiques. Les contextes lexico-syntaxiques des termes en relation sont comparés deux à deux et leurs scores de contiguïté forment une matrice carrée. Chaque composante connexe décrite dans la matrice reflète une classe au sein de laquelle les contextes lexico-syntaxiques sont proches. Les patrons sont alors définis par abstraction des contextes à partir des composantes connexes puis appliqués sur le corpus pour extraire de nouvelles relations. Ce processus itératif qui, après une validation manuelle, exploite les relations acquises automatiquement pour acquérir de nouveaux patrons, permet d'obtenir une F1-mesure moyenne de 0,53 sur des relations d'hyperonymie avec 836 couples de termes. De la même manière, la méthode proposée dans DIPRE (Brin, 1999) s'appuie sur les instances de mots impliqués dans une relation pour induire un patron. Un classifieur est ensuite appliqué pour extraire de nouvelles relations à partir du patron obtenu et ainsi obtenir d'autres exemples qui seront ajoutés à la base d'apprentissage. L'architecture de DIPRE est exploité dans Snowball (Agichtein & Gravano, 2000) pour identifier les relations entre entités nommées (organisation,lieu) sur des textes de langue générale. Cette méthode est plus flexible que la précédente : elle ne vise pas à trouver des correspondances exactes mais à identifier des similarités en tenant compte de légères variations au niveau des mots ou de la ponctuation. La F1-mesure de DIPRE est de 0,68 contre 0,87 pour Snowball.

Ces approches permettent de bien appréhender les mentions de relations. Toutefois, les patrons obtenus ne peuvent pas toujours prendre en compte les variations ou l'inclusion de mots supplémentaires. Il est possible de remédier à ce problème en s'appuyant sur un alignement séquentiel multiple pour identifier les contextes similaires (Meng & Morioka, 2015). Les points communs et les disparités entre les groupes de phrases sont pris en compte lors de l'alignement et permettent d'identifier systématiquement les zones de similarité et de variabilité dans les contextes des termes en relation. L'approche appliquée sur un corpus biomédical obtient une F1-mesure moyenne de 0,94 sur des tâches de localisation de maladie et d'identification de valeurs numériques (taille ou date) à partir de 7 365 rapports de radiologie. L'extraction de patrons linguistiques peut également être basée sur des séquences d'*items* (Cellier *et al.*, 2010). Chaque mot est décrit par un ensemble d'informations (sa forme fléchie, son lemme et sa catégorie grammaticale) et est représenté sous forme d'ensemble de descripteurs. Il est alors possible d'obtenir des patrons lexico-syntaxique tenant compte de l'ordre partiel des motifs et en ne retenant que les plus fréquents. L'application de cette méthode sur un corpus médical permet d'obtenir une F1-mesure moyenne de 0,29 (Holat *et al.*, 2016) sur des tâches d'extraction de symptôme à partir de 10 000 résumés Medline. Une approche basée sur des patrons verbaux a également été proposée (Song *et al.*, 2015). Pour extraire des relations, les verbes d'une phrase situés entre deux entités sont identifiés et comparés à une liste de verbes à caractéristiques relationnels. La nominalisation du verbe est également prise en compte. Les règles d'extraction des relations reposent principalement sur l'analyse syntaxique : étant donné deux entités, l'arbre syntaxique indique les dépendances syntaxiques entre eux. Pour chaque verbe dans un chemin de dépendance, il existe un chemin à gauche du verbe (lemmatisé) et à droite de la forme $sujet \rightarrow interagit \leftarrow prep_in \leftarrow modifieur$. Les nominalisations sont identifiées par des modèles de la forme `<NOMINALIZATION_TERM><PREP (POS)> <ENTITE_A> <PREP (POS)> <ENTITE_B>`. L'approche permet d'avoir une F1-mesure autour de 0,80 sur des corpus médicaux sur des tâches

d'extraction de relation entre protéines, ou entre gène et cancer à partir de 225 résumés Medline.

Classification supervisée Les méthodes d'apprentissage supervisé sont de plus en plus souvent mises en œuvre pour extraire de relations sémantiques à partir de corpus annotés. Kambhatla (2004) utilise plusieurs types de descripteurs pour décrire les exemples de relations et pour entraîner un classifieur log-linéaire. La meilleure F1-mesure (0,409) est obtenue avec les descripteurs constitués d'arbres d'analyse syntaxique sur des tâches d'extraction d'emplacement ou de rôle dans des textes généraux avec 9 752 instances de relations. A la différence des méthodes classiques, les descripteurs ne sont pas explicitement générés pour les méthodes à base de noyau : les exemples conservent leurs représentations originales (des analyses superficielles) et sont utilisés dans les algorithmes d'apprentissage uniquement en calculant une fonction de similarité (ou de noyau) entre eux (Zelenko *et al.*, 2003). Zelenko *et al.* (2003) ont évalué un Perceptron et un classifieur SVM utilisant soit des noyaux de sous-arbres contigus, soit des noyaux de sous-arbres éparses. L'approche permet d'obtenir une F-mesure de 0,86 pour la relation *Personne-Affiliation* et 0,83 pour la relation *Organisation-Lieu*. Dans le domaine médical, de nombreux travaux ont pour objectif l'extraction des interactions entre médicaments (*Drug-Drug Interactions* - DDI) par méthode d'apprentissage. Ben Abacha *et al.* (2015) se basent sur une méthode hybride basée sur un classifieur SVM pour extraire des relations en combinant : (i) une méthode d'apprentissage automatique avec comme descripteurs les mots dans le contexte des exemples, (ii) une méthode à base de noyau utilisant les arbres de dépendance. L'union et l'intersection des résultats de chaque méthode permettent d'obtenir des F1-mesures de 0,5 et 0,39 respectivement. Ben Abacha *et al.* (2015) ont également proposé une deuxième approche permettant d'extraire les interactions entre médicaments en deux étapes : (i) la détection des DDIs potentiels et (ii) la classification des relations précédemment identifiées à l'aide d'un classifieur SVM linéaire. Les F1-mesures pour la détection de DDI seule et pour la détection et la classification sont respectivement 0,53 et 0,40 sur des résumés Medline et 0,83 et 0,68 sur des documents issues de DrugBank.

Cejuela *et al.* (2018) considèrent l'extraction de relation de localisation de protéine comme une classification binaire. Tous les couples protéine-localisation apparaissant dans une même phrase sont extraits pour former des instances positives s'ils sont en relation, et négatives sinon. Plusieurs types de descripteurs ont été envisagés : le nombre d'entités (protéine ou localisation) dans une phrase, les n-grammes des éléments de texte présents entre les deux entités, les n-grammes des mots présents dans l'arbre de dépendance permettant de relier les deux entités. Pour les descripteurs à base de n-grammes, les mots sont représentés par leurs lemmes, leurs étiquettes morpho-syntaxiques ou les nœuds de dépendances syntaxiques (par exemple préposition ou objet direct). Un sélecteur de descripteurs non supervisé a ensuite été appliqué sur l'ensemble pour ne retenir que les plus efficaces. Enfin, un classifieur SVM a été entraîné sur les instances et permet d'obtenir une F1-mesure de 0,74 sur des résumés Medline.

Apprentissage profond Les réseaux de neurones profonds (Deep Neural Network - DNN) ont récemment montré un grand potentiel pour de nombreuses tâches de TAL y compris l'extraction de relations. Zeng *et al.* (2014) ont proposé une méthode basée sur une architecture à quatre étapes : (1) les mots sont transformés en vecteurs en recherchant les plongements de mots ; (2) des descripteurs tiennent compte du lexique (les deux entités en relations, les mots à gauche et à droite, les hyperonymes issus de WordNet) et au niveau des phrases (le plongements des mots et des positions par rapport aux entités en relation) ; (3) les vecteurs obtenus sont directement concaténés pour former le vecteur final ; (4) pour calculer la confiance de chaque relation, le vecteur de caractéristiques est introduit

dans un classifieur *softmax*. La sortie du classifieur est un vecteur dont la dimension est égale au nombre de types de relation prédéfinis. La valeur de chaque dimension est le score de confiance de la relation. L'approche a une F1-mesure de 0,827. Nguyen & Grishman (2015) ont proposé une approche rajoutant plusieurs fenêtres de diverses tailles pour définir les filtres convolutifs. Cette stratégie permet au réseau de capturer des plages de n-grammes plus larges et faciliter l'extraction des relations. L'initialisation de la méthode est réalisée à l'aide de plongements de mots pré-entraînés. Les plongements de mots et de position sont ensuite optimisés comme paramètres du modèle. Le modèle est alors entraîné sans ressource ou annotation supplémentaire. La meilleure F1-mesure de 0,6132 est obtenue avec une combinaison de fenêtres dont la taille varie entre 2 et 5. L'extraction d'interaction entre médicaments peut également s'appuyer sur un modèle de réseau de neurone convolutif (CNN) (Liu *et al.*, 2016). Chaque paire de médicaments est classée dans l'un des types prédéfinis de DDI (*Mechanism*/*mécanisme*, *Effect*/*effet*, *Advice*/*recommandation*, *Int*/*existence d'une interaction*) ou classée comme une paire de médicaments n'interagissant pas. Pour assurer la généralisation des méthodes basées sur l'apprentissage, les deux médicaments en question sont normalisés et remplacés par "drug1" et "drug2" dans leur ordre d'apparition et les autres médicaments sont remplacés par "drug0". Les phrases sont segmentées en mots et converties en minuscules, puis les instances négatives sont éliminées. En plus des plongements de mots, les plongements de distance sont intégrées au modèle pour encoder les distances relatives entre les mots et les deux médicaments en question. Les F1-mesures sur des résumés Medline et des documents issus de Drugbank sont respectivement 0,5212 et 0,7152.

Co-occurrences Alors que les méthodes présentées ci-dessus tentent d'extraire les relations au niveau des phrases seules, des méthodes visent à identifier les entités dont la co-occurrence est statistiquement significative. Ainsi, Ramani *et al.* (2005) identifient les protéines en interaction en comparant le nombre de résumés citant une paire de protéines avec la probabilité de co-citation sous un modèle aléatoire. Un classifieur bayésien basé sur la fréquence d'utilisation des mots pertinents pour les interactions de protéines est ensuite appliqué. Cette méthode peut également être combinée avec une extraction dans les phrases individuelles en remplaçant le nombre de co-citation de deux protéines par la somme des scores de précision de chaque co-citation (Bunescu *et al.*, 2006). La qualité des résultats augmente alors de plus de 40%. Dans la même optique, Lee *et al.* (2018) proposent une méthode d'extraction de relations à la fois au niveau du document entier et au niveau d'une phrase seule. D'une part, toutes les combinaisons possibles de relations au niveau du document sont considérées comme correctes si la relation est vraie dans n'importe quelle partie du document. D'autre part, au niveau de la phrase, seule la co-occurrence des entités est utilisée sans considérer la fréquence des entités ou leur contexte. Les relations extraites sont ensuite décrites à l'aide de la distance entre les entités, de la fréquence des entités, et en fonction des scores obtenu par des requêtes dans BEST (Biomedical entity search tool) (Lee *et al.*, 2016). Les instances des relations sont alors vectorisées par méthode de plongement de mots à 300 dimensions avant d'être injectées dans des classifieurs à base d'arbre de décision, de forêt aléatoire (*RandomForest*) ou de réseaux de neurones convolutifs en équilibrant à chaque fois le nombre d'instances positives et négatives. L'extraction au niveau du document permet d'obtenir des F1-mesures de 0,958 pour la relation *mutation-gène* et 0,821 pour la relation *mutation-médicament*. L'extraction au niveau des phrases permet d'obtenir des F1-mesures de 0,955 pour la relation *mutation-gène* et 0,856 pour la relation *mutation-médicament*.

Dans cet article, nos expériences se basent sur la méthode d'extraction dans une phrase seule comme proposée par Lee *et al.* (2018) en utilisant l'approche de Cejuela *et al.* (2018) pour former les instances : tous les couples *aliment-médicament* , *complément_alimentaire-médicament* ou *aliment-*

effet_secondaire apparaissant dans une même phrase sont extraits pour former des instances positives s'ils sont en relation et négatives sinon. Ensuite, nous définissons une méthode à deux étapes : détection et classification comme l'a proposé par Ben Abacha *et al.* (2015). Comme dans Zelenko *et al.* (2003), nous n'avons pas explicitement introduit des descripteurs. En revanche, nous avons suivi la logique de Liu *et al.* (2016) pour généraliser les entités nommées. Parmi les cinq classifieurs que nous avons expérimentés, quatre sont mentionnés dans l'état de l'art : un SVM linéaire et un arbre de décision (Cejuela *et al.*, 2018), un classifieur bayésien (Ramani *et al.*, 2005), un perceptron multi-couches (Zelenko *et al.*, 2003).

3 Corpus

Notre corpus de travail est constitué de 639 résumés d'articles scientifiques du domaine médical. Ils ont été collectés sur le portail PubMed [3] grâce à la requête :

```
(FOOD DRUG INTERACTIONS"[MH] OR "FOOD DRUG INTERACTIONS*" )
                              AND ("adverse effects*")
```

L'ensemble des 639 résumés a été annoté sous Brat (Stenetorp *et al.*, 2012) par un externe en pharmacie, en se focalisant sur les informations relatives à la relation entre les aliments, les médicaments et les pathologies. Neuf types de termes et vingt-et-un types de relations ont été extraits de ces annotations. Les neuf types d'entités annotés font références aux aliments et leurs composants, aux médicaments et leur composant ainsi que diverses informations relatives à la prise de médicament et à la consommation d'aliment. Les catégories sémantiques sont : `Drug`, `Food`, `MealTime`, `DrugEffect`, `Treated disease`, `Side effect`, `Numbers`, `Frequency`, `Dosage`, `Food supplement`. Les vingt-et-un types de relations décrivent : (1) la diminution ou suppression d'un effet de médicament due à un aliment ; (2) l'augmentation ou la diminution de l'effet d'un médicament ; (3) l'apparition de nouveaux effets indésirables encore inconnus pour un médicament ; (4) existence de relation entre médicament et aliment sans plus de précision. Ces annotations seront nos références lors de l'évaluation de notre méthode. Le processus de collecte du corpus et le schéma d'annotations sont détaillés dans (Hamon *et al.*, 2017). La figure 1 présente un extrait de ces annotations. Nous disposons de 2 341 instances positives et 25 231 instances négatives de relations.

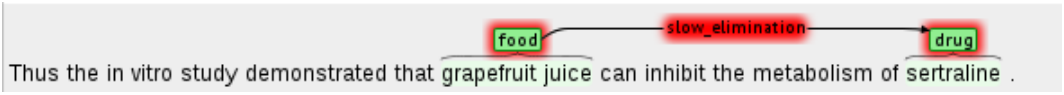

FIGURE 1 – Exemple d'annotation Brat d'un résumé Medline

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

4 Méthode

Nous considérons l'identification des relations entre aliment et médicament comme un problème de classification supervisée. Toutefois, à travers l'analyse des annotations, nous pouvons constater que l'ensemble d'apprentissage est très déséquilibré (peu d'exemples positifs et beaucoup d'exemples négatifs) et les types de relations recherchés ont une granularité assez fine. Aussi, nous proposons une méthode réalisant une classification des exemples suivant quatre étapes : (1) un classifieur binaire a pour objectif de sélectionner les phrases des résumés contenant les relations pertinentes ; (2) un second classifieur identifie les trois classes regroupant les types de relation ; (3) un troisième classifieur catégorise les relations correspondant à des interactions entre aliment et médicament (4) un dernier classifieur reconnaît les entités (aliment et médicament) en interaction dans les phrases sélectionnées lors de l'étape précédente. Actuellement, nous nous intéressons au deux premières étapes.

Pour cela, nous définissons une représentation vectorielle et plusieurs ensembles de descripteurs. Puis, nous avons évalué les performances de plusieurs algorithmes de classification proposés dans la boîte à outil Scikit-Learn [4].

4.1 Vectorisation

Pour pouvoir entraîner les classifieurs, les phrases sont représentées sous forme de vecteurs dans un espace vectoriel à 6 805 dimensions correspondant aux mots du dictionnaire construit à partir des résumés de notre corpus. Nous utilisons une vectorisation à base de comptage de mot : chaque phrase est représentée par le vecteur correspondant aux nombres d'occurrences de chaque mot de l'ensemble du vocabulaire dans la phrase en question.

4.2 Classification et descripteurs

Etape 1 : Classification binaire des phrases. La première étape de notre approche consiste à éliminer les instances négatives c'est-à-dire les phrases contenant des paires de médicament et aliment n'impliquant pas d'interaction. Pour cela, nous entraînons un classifieur qui détecte l'existence d'une interaction entre médicament et aliment pour chaque phrase contenant une paire médicament-aliment. Les instances sont classées comme positives si elle implique une relation entre le médicament et l'aliment en question, et négatives sinon.

Etape 2 : Classification multi-classe regroupant les types de relation. Pour avoir plus de précision sur les types de relation, la deuxième étape de notre approche consiste en une classification multi-classes sur les instances étiquetées positives déduites de l'étape 1. Selon les schémas d'annotation, nous avons défini trois classes de relation : (1) les relations impliquant directement un aliment et un médicament (augmentation ou diminution de l'absorption d'un médicament, augmentation ou diminution de la vitesse d'absorption ou d'élimination d'un médicament, effet positif ou négatif sur un médicament, etc.) représentées par 352 exemples, (2) les relations impliquant l'effet d'un médicament et un problème médical (augmentation d'un effet indésirable d'un médicament, apparition d'un nouvel effet indésirable, etc.) représentées par 1260 exemples, (3) la présence d'une relation impliquant un

4. http://scikit-learn.org/stable/

aliment et un médicament sans plus de précision (729 exemples). Chaque instance est étiquetée par la classe de relation qu'elle implique.

Algorithmes. Nous avons comparé les performances de cinq algorithmes de classification en gardant les paramètres par défaut fournis par Scikit-Learn : (1) un classifieur à base d'arbre de décision (DecisionTree), (2) un classifieur SVM linéaire(linearSVC), (3) un classifieur bayésien naïf multinomial (MultinomialNB), (4) un classifieur à base de régression (LogisticRegression), (5) un classifieur Perceptron à quatre couches (MLP).

Description des exemples d'apprentissage. Nous avons choisi quatre ensembles de descripteurs sur lesquels sont entraînés les modèles :

1. Mots sous leur forme fléchie

   *ex : Bioavailability enhancement by **grapefruit juice** noted with other **dihydropyridine calcium antagonists** does not occur with amlodipine.*

2. Mots sous leur forme fléchie et termes suivis de leur catégorie sémantique

   *ex : Bioavailability enhancement by **grapefruit juice /food/** noted with other **dihydropyridine calcium antagonists /drug/** does not occur with amlodipine.*

3. Termes remplacés par leur catégorie sémantique pour généraliser le texte sans perdre l'information concernant la nature des entités

   *ex : Bioavailability enhancement by **food** noted with other **drug** does not occur with amlodipine.*

4. Normalisation des arguments des relations (`arg1` et `arg2`) pour ne permettre aucune différenciation entre les entités d'une relation candidate

   *ex : Bioavailability enhancement by **arg1** noted with other **arg2** does not occur with amlodipine.*

4.3 Evaluation

Pour chaque ensemble de descripteurs, nous avons effectué deux évaluations : (1) une partie (67%) du corpus est utilisée pour l'entraînement tant que l'autre (33%), le corpus de test, est utilisée pour l'évaluation des modèles ; (2) une validation croisée en dix échantillons des modèles. Nous avons calculé la précision, le rappel, la F1-mesure. La précision est la portion de réponses correctes parmi les réponses données. Le rappel est la portion de réponses correctes parmi les réponses attendues. La F1-mesure est la moyenne harmonique de la précision et du rappel. Nous avons effectué une macro-évaluation de nos résultats (Sebastiani, 2002). La macro-évaluation est la moyenne des évaluations (précision, rappel) obtenues pour chaque classe. Dans notre cas où les classes sont très déséquilibrées, elle fournit une information plus fiable qu'une évaluation classique calculée sur les résultats indépendamment des classes. Nous avons également évalué les moyennes et écart-types des résultats de la validation croisée afin d'identifier leurs disparités.

5 Résultats

Nous présentons les résultats obtenus pour la sélection des phrases puis la catégorisation par type de relation.

Etape 1 : Classification binaire des phrases. Les tables 1, 2, 3 et 4 présentent les résultats obtenus pour identifier les phrases contenant des relations pertinentes. Suivant les descripteurs utilisés, la F-mesure varie entre 0,54 et 0,71. Nous observons que les modèles appris sur les descripteurs correspondant aux formes fléchies des mots pour décrire les exemples, ne permettent pas d'obtenir des résultats concluants (table 1). L'utilisation des descripteurs génériques `arg1` et `arg2` a un impact positif. Les meilleurs résultats sont obtenus avec les arbres de décision et le Perceptron (MLP) en utilisant les catégories sémantiques des termes avec et sans les mots (tables 2 et 3). Alors que les modèles tendent à équilibrer le rappel et la précision ou à favoriser la précision, les classifieurs bayésien naïfs (MultinomialNB) ont tendance à privilégier le rappel par rapport à la précision. Enfin, à l'exception de l'utilisation du Perceptron (MLP) sur l'ensemble des descripteurs combinant les mots et les catégories de termes (table 3), les évaluations obtenues par validation croisée et sur le corpus de test sont similaires.

Modèle	Test			Cross-Validation	
	F1	P	R	F1	Ecart-type
DecisonTree	0,55	0,63	0,54	0,53	$\pm$ 0,012
LinearSVC	0,56	0,63	0,55	0,56	$\pm$ 0,019
MultinomialNB	0,55	0,56	0,68	0,55	$\pm$ 0,011
LogisticRegression	0,54	0,62	0,53	0,53	$\pm$ 0,013
MLP	**0,59**	0,64	0,57	0,57	$\pm$ 0,033

TABLE 1 – Sélection des phrases contenant les relations (descripteurs : formes fléchies des mots)

Algo	Test			Cross-Validation	
	F1	P	R	F1	Ecart-type
DecisonTree	**0,70**	0,72	0,68	0,68	$\pm$ 0,023
LinearSVC	0,59	0,64	0,58	0,58	$\pm$ 0,022
MultinomialNB	0,56	0,57	0,70	0,55	$\pm$ 0,011
LogisticRegression	0,56	0,64	0,55	0,56	$\pm$ 0,015
MLP	**0,70**	0,75	0,67	0,61	$\pm$ 0,111

TABLE 2 – Sélection des phrases contenant les relations (descripteurs : mots et catégories sémantiques des termes)

Etape 2 : Classification multi-classe regroupant les types de relation. Les tables 5, 6, 7 et 8 présentent les résultats obtenus lors de la reconnaissance des regroupements de types de relations. L'utilisation des mots mais aussi des termes normalisés (tables 5 et 8) conduisent à des résultats beaucoup plus faibles que les deux autres ensembles de descripteurs (tables 6 et 7). Comme pour l'étape précédente, l'utilisation des catégories sémantiques des termes a un impact positif sur les résultats. Parmi les modèles utilisés, le classifieur bayésien naïf (`MultinomialNB`) est celui conduisant

Algo	Test			Cross-Validation	
	F1	P	R	F1	Ecart-type
DecisonTree	0,70	0,71	0,69	0,69	± 0,019
LinearSVC	0,64	0,67	0,62	0,64	± 0,022
MultinomialNB	0,57	0,57	0,68	0,57	± 0,011
LogisticRegression	0,59	0,69	0,57	0,58	± 0,014
MLP	**0,71**	0,73	0,69	0,72	± 0,017

TABLE 3 – Sélection des phrases contenant les relations (descripteurs : catégories sémantiques des termes)

Algo	Test			Cross-Validation	
	F1	P	R	F1	Ecart-type
DecisonTree	0,62	0,62	0,61	0,63	± 0,018
LinearSVC	0,64	0,67	0,62	0,64	± 0,017
MultinomialNB	0,56	0,57	0,66	0,56	± 0,014
LogisticRegression	0,59	0,68	0,57	0,58	± 0,013
MLP	**0,67**	0,67	0,66	0,67	± 0,014

TABLE 4 – Sélection des phrases contenant les relations (descripteurs : termes normalisés, remplacés par `arg1` et `arg2`)

aux résultats les plus faibles. En revanche, nous observons que pour cette tâche de classification, la régression logistique (`LogisticRegression`) mais aussi l'arbre de décision (`DecisonTree`) et le classifieur SVM (`linearSVC`) permettent d'obtenir de bons résultats lorsqu'ils sont utilisés sur des exemples décrits par les formes fléchies des mots et les catégories sémantiques des termes.

Algo	Test			Cross-Validation	
	F1	P	R	F1	Ecart-type
DecisonTree	0,60	0,60	0,60	0,59	± 0,031
LinearSVC	**0,65**	0,65	0,65	0,66	± 0,036
MultinomialNB	0,63	0,62	0,64	0,65	± 0,037
LogisticRegression	**0,65**	0,66	0,65	0,66	± 0,029
MLP	**0,65**	0,65	0,64	0,63	± 0,058

TABLE 5 – Classification des regroupements des types de relation (descripteurs : formes fléchies des mots)

Discussion Alors que nous avons utilisé des ensembles de descripteurs issus de l'état de l'art, ces premiers résultats nous permettent de mieux cibler les méthodes à développer pour identifier les relations entre un médicament et un aliment. Ainsi, les catégories sémantiques des termes associées aux formes fléchies des mots du corpus contribuent à obtenir des résultats intéressants à la fois pour sélectionner les phrases contenant des relations pertinentes et pour identifier les grandes classes de relations que nous cherchons à identifier. Suivant les étapes de notre méthode, nous avons pu constater que les résultats variaient suivant les classifieurs utilisés. Ainsi, pour sélectionner les phrases intéressantes, il est préférable d'utiliser un classifieur Perceptron ou un arbre de décision. En revanche,

Algo	Test			Cross-Validation	
	F1	P	R	F1	Ecart-type
DecisonTree	**0,78**	0,79	0,78	0,74	± 0,020
LinearSVC	**0,78**	0,79	0,77	0,78	± 0,021
MultinomialNB	0,67	0,67	0,68	0,70	± 0,039
LogisticRegression	**0,78**	0,79	0,76	0,77	± 0,026
MLP	0,73	0,75	0,72	0,67	± 0,147

TABLE 6 – Classification des regroupements des types de relation (descripteurs : mots et catégories sémantiques des termes)

Algo	Test			Cross-Validation	
	F1	P	R	F1	Ecart-type
DecisonTree	0,76	0,76	0,76	0,71	± 0,036
LinearSVC	0,77	0,77	0,77	0,78	± 0,017
MultinomialNB	0,69	0,69	0,69	0,69	± 0,041
LogisticRegression	**0,78**	0,79	0,78	0,77	± 0,026
MLP	0,73	0,72	0,74	0,62	± 0,166

TABLE 7 – Classification des regroupements des types de relation (descripteurs : catégories sémantiques des termes)

Algo	Test			Cross-Validation	
	F1	P	R	F1	Ecart-type
DecisonTree	0,61	0,61	0,61	0,60	± 0,027
LinearSVC	**0,66**	0,66	0,66	0,64	± 0,028
MultinomialNB	0,62	0,62	0,62	0,61	± 0,039
LogisticRegression	**0,66**	0,67	0,66	0,66	± 0,023
MLP	0,64	0,64	0,64	0,53	± 0,128

TABLE 8 – Classification des regroupements des types de relation (descripteurs : termes normalisés, remplacés par `arg1` et `arg2`)

pour reconnaître les regroupements de types de relations, une régression logistique, un classifieur SVM ou un arbre de décision sont plus adaptés.

6 Conclusion et perspectives

Notre article propose un premier pas vers l'extraction d'interaction entre médicament et aliment. De très nombreuses approches ont été proposées pour l'extraction de relation entre entités dans les textes de spécialité. Il peut s'agir de méthodes utilisant des patrons décrivant le contexte caractéristique des entités en relation mais aussi de méthodes de classification supervisée s'appuyant sur le noyau d'un classifieur ou sur des ensembles de descripteurs. Plus récemment, l'extraction de relations peut être réalisée grâce à un apprentissage profond et des plongements de mots et de distances. Les méthodes d'extraction de relations sur les textes biomédicaux s'intéressent à de nombreux types de relations entre des entités (interaction entre médicaments, relations entre médicaments et maladie, etc.).

Dans ce travail, notre objectif est d'identifier des relations entre médicament et aliment. Pour cela, nous avons proposé une méthode en quatre étapes visant à cibler précisément les informations qui nous intéressent. Pour l'instant, les deux premières étapes (sélection des phrases contenant les relations et identification des regroupements de types de relations) ont été mises en œuvre sous forme d'un apprentissage supervisé. Nous nous sommes intéressés au choix des classifieurs à utiliser mais aussi aux descripteurs nécessaires pour l'apprentissage supervisé. A travers une première série d'expériences, nous avons pu observer que les descripteurs incluant les catégories sémantiques des termes permettent d'obtenir les meilleurs résultats. Aussi, le classifieur le plus performant varie suivant l'étape : un Perceptron multi-couche est le plus adapté pour la tâche de sélection des phrases contenant les relations, tandis que pour la détection des grandes classes de relations, une régression logistique ou un SVM linéaire est plus adapté.

Dans la suite de ce premier travail, nous allons nous intéresser d'une part, à l'étape suivante de notre approche, c'est-à-dire la reconnaissance des différents types de relations et d'autre part, à l'identification des entités en relation (aliment, médicament, maladie, etc). Pour cela, nous allons nous appuyer sur des méthodes de classification mais aussi à l'utilisation de ressources terminologiques. Les résultats préliminaires présentés dans cet article doivent être améliorés. Nous envisageons l'utilisation d'autres méthodes de classification comme les réseaux de neurones profonds entièrement connectés combinés à des plongements des mots. Nous souhaitons également étudier l'impact d'autres descripteurs (lemmes des mots, catégories morpho-syntaxiques, relations syntaxiques, catégories sémantiques des termes avec différents niveaux de granularité, etc.) mais aussi des méthodes d'échantillonnage afin de réduire le déséquilibre des données.

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.

AGICHTEIN E. & GRAVANO L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries (DL'00)*, p. 85–94, New York, NY, USA : ACM.

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.

BRIN S. (1999). *Extracting Patterns and Relations from the World Wide Web.* Technical Report 1999-65, Stanford InfoLab.

BUNESCU R., MOONEY R., RAMANI A. & MARCOTTE E. (2006). Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline. In

Proceedings of the Workshop on Linking Natural Language Processing and Biology (BioNLP '06), p. 49–56, Stroudsburg, PA, USA : Association for Computational Linguistics.

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.

CELLIER P., CHARNOIS T. & PLANTEVIT M. (2010). Sequential patterns to discover and characterise biological relations. 11th international conference on intelligent text processing and computational linguistics. *Proceedings of CICLing'10*, **Springer-Verlag**(6008), 537–548.

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.

DOOGUE M. & POLASEK T. (2013). The abcd of clinical pharmacokinetics. *Ther Adv Drug Saf*, **4**(1), 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.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, p. 539–545 : Association for Computational Linguistics.

HOLAT P., TOMEH N., CHARNOIS T., BATTISTELLI D., JAULENT M.-C. & MÉTIVIER J.-P. (2016). Fouille de motifs et crf pour la reconnaissance de symptômes dans les textes biomédicaux. In *Actes de la conférence JEP-TALN-RECITAL 2016*, volume 2, p. 194–206, Paris, France.

KAMBHATLA N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, p. 178–181, Barcelona, Spain : Association for Computational Linguistics.

LEE K., KIM B., CHOI Y., KIM S., SHIN W., LEE S., PARK S., KIM S., TAN A. C. & KANG J. (2018). Deep learning of mutation-gene-drug relations from the literature. *BMC Bioinformatics*, **19**(1), 21.

LEE S., KIM D., LEE K., CHOI J., KIM S., JEON M., LIM S., CHOI D., KIM S., TAN A.-C. & KANG J. (2016). BEST : Next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLoS One*, **11**(10), e0164680.

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.

MORIN E. (1998). Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes. In *Actes de la 5ème conférence sur le Traitement Automatique des Langues Naturelles*, p. 172–181, Paris, France : Association pour le Traitement Automatique des Langues.

NGUYEN T. H. & GRISHMAN R. (2015). Relation extraction : Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, p. 39–48 : Association for Computational Linguistics.

RAMANI A. K., BUNESCU R. C., MOONEY R. J. & MARCOTTE E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*, **6**(5), R40–R40.

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

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

SONG M., CHUL KIM W., LEE D., EUN HEO G. & YOUNG KANG K. (2015). PKDE4J : Entity and relation extraction for public knowledge discovery. *Journal of Biomedical Informatics*, **57**, 320–332.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, p. 102–107, Stroudsburg, PA, USA : Association for Computational Linguistics.

ZELENKO D., AONE C. & RICHARDELLA A. (2003). Kernel methods for relation extraction. *J. Mach. Learn. Res.*, **3**, 1083–1106.

ZENG D., LIU K., LAI S., ZHOU G. & ZHAO J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, p. 2335–2344.