Extraction automatique des interactions aliment-médicament à partir d'articles scientifiques

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

RÉSUMÉ
Dans cet article, nous nous intéressons à l'extraction des interactions aliment-médicament, une tâche similaire à l'extraction de la relation entre les termes dans des textes spécialisés. Nous présentons une méthode d'apprentissage supervisé et les résultats d'une première série d'expériences. Malgré le déséquilibre des classes, les résultats sont encourageants. Nous avons identifié les classifieurs les plus pertinents en fonction des étapes de notre méthode. Nous avons également observé l'impact important des catégories sémantiques des termes utilisés comme caractéristiques.

MOTS-CLÉS : Interaction aliment-médicament, Relation sémantique, Corpus spécialisés, Apprentissage supervisé.

KEYWORDS: Food-drug interaction, Semantic relation, Specialized corpora, Supervised Learning.

1 Introduction

Bien que des bases de connaissances ou des terminologies existent dans des domaines, la mise à jour de ces informations nécessite souvent des données non structurées telles que les articles scientifiques. Ce constat est d'autant plus vrai lorsque les connaissances à recenser ne sont pas déjà présentes dans une base. Ainsi, alors que les interactions entre médicaments (Aagaard & Hansen, 2013) ou les effets indésirables des médicaments (Aronson & Ferner, 2005) sont répertoriés dans des bases de données telles que DrugBank [1] ou Theriaque [2], d'autres informations telles que les interactions aliment-médicament sont à peine répertoriées dans les bases de connaissances, dispersées dans la littérature scientifique principalement sous forme textuelle. Cependant, les interactions aliment-médicament correspondent à divers types d'effets indésirables des médicaments et entraînent des conséquences néfastes sur la santé et le bien-être du patient. Dans cet article [3], nous nous intéressons à l'extraction de ces interactions en tant que tâche d'acquisition de relations, étant donné les références à un aliment et à un médicament dans les résumés d'articles scientifiques.

Notre contribution est : (1) la sélection de phrases pertinentes pour l'interaction aliment-médicament, (2) la classification des relations dans les phrases pertinentes.

2 Expériences

2.1 Corpus

Notre corpus est constitué de 2 341 occurrences positives et de 25 231 occurrences négatives extraites de 639 résumés d'articles scientifiques rassemblés sur le portail PubMed à l'aide de la requête suivante : (FOOD DRUG INTERACTIONS"[MH] OR "FOOD DRUG INTERACTIONS*") AND ("adverse effects*") annotés avec `Brat`. Le processus de collecte de corpus et le schéma d'annotation sont détaillés dans (Hamon *et al.*, 2017). Les instances positives sont classées en 21 types de relations mais nous avons regroupé ces relations en 3 groupes : interaction directe aliment-médicament (352 relations), effet indésirable lié au médicament (1 260 relations), relation sans précision (729 relations). Nous avons ensuite vectorisé ces phrases en fonction du nombre de mots : chaque phrase est représentée par un vecteur correspondant au nombre d'occurrences de chaque mot du vocabulaire dans la phrase.

1. https://www.drugbank.ca/
2. http://www.theriaque.org
3. Ce travail a été soutenu par l'ANR via la subvention ANR-16-CE23-0012 (projet MIAM).

2.2 Classification et descripteurs

Algorithmes. Nous comparons les performances de cinq algorithmes de classification avec les paramètres par défaut fournis par Scikit-Learn [4] : (1) un arbre de décision (DecTree), (2) un classifieur SVM linéaire (LSVC), (3) un classifieur Naive Bayes multinomial (MNB), (4) une régression logistique (LogReg), (5) un Perceptron à quatre couches (MLP). Nous évaluons nos modèles en utilisant la F1-mesure calculée sur une validation croisée de 10 échantillons. La F1-mesure (F1) est la moyenne harmonique de la précision (P) et du rappel (R) tels que

$$P = \frac{nb\ réponses\ correctes}{nb\ réponses\ retournées} \qquad R = \frac{nb\ réponses\ correctes}{nb\ réponses\ attendues} \qquad F1 = 2.\frac{P * R}{P + R}$$

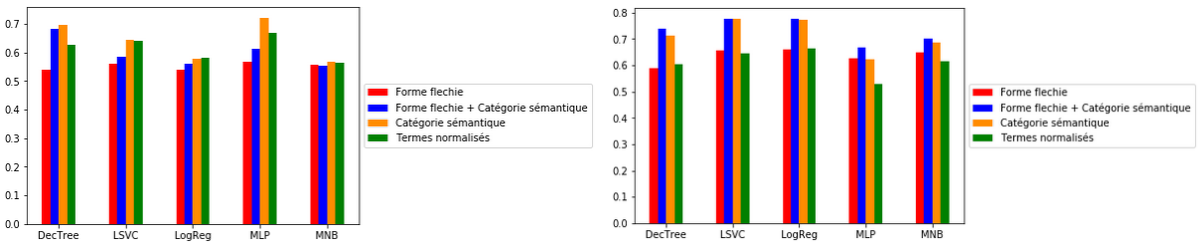
Pour entraîner nos modèles, nous avons utilisé quatre types de descripteurs :

1. Forme fléchie des mots (*ex : Bioavailability enhancement by **grapefruit juice** noted with **dihydropyridine calcium antagonists** does not occur with amlodipine.*)
2. Forme fléchie et catégorie sémantique des termes (*ex : Bioavailability enhancement by **grapefruit juice /food/** noted with **dihydropyridine calcium antagonists /drug/** does not occur with amlodipine.*)
3. Termes remplacés par leur catégorie sémantique (*ex : Bioavailability enhancement by **food** noted with **drug** does not occur with amlodipine.*)
4. Forme fléchie et terme normalisé (*ex : Bioavailability enhancement by **arg1** noted with **arg2** does not occur with amlodipine.*)

3 Résultats

Etape 1 : Classification binaire. La figure 1a présente les résultats obtenus pour identifier les phrases contenant des relations pertinentes. Selon les descripteurs utilisés, la F1-mesure varie entre 0,54 et 0,71. Les meilleurs résultats sont obtenus avec un arbre de décision et un perceptron en utilisant les catégories sémantiques.

Etape 2 : Classification multiclasse. La figure 1b présente les résultats obtenus lors de la reconnaissance des groupes de relations. Comme à l'étape précédente, l'utilisation des catégories sémantiques de termes a un impact positif sur les résultats. Parmi les modèles utilisés, le classifieur Naive Bayes est celui qui conduit aux résultats les plus faibles. D'autre part, nous obtenons de bons résultats avec une régression logistique, un arbre de décision et un SVM linéaire.



(a) Etape 1 : Classification Binaire

(b) Etape 2 : Classification multiclasse

FIGURE 1: F1-mesure sur 10 échantillons

4 Conclusion et perspectives

Nous proposons un premier pas vers l'extraction des interactions aliment-médicament. Ces premières expériences montrent que des descripteurs telles que les catégories sémantiques donnent de meilleurs résultats. Comme perspectives, nous allons poursuivre en réalisant les deux prochaines étapes de notre méthode : (1) la reconnaissance fine des différents types de relations et (2) l'identification des entités en interaction (aliments, médicaments, maladies, etc.). Les résultats préliminaires présentés dans cet article doivent être améliorés. Nous envisageons d'utiliser d'autres méthodes de classification telles que les réseaux de neurones profonds convolutionnels utilisant des plongements de mots. Nous souhaitons également étudier l'impact d'autres descripteurs (lemme, catégorie morpho-syntaxique, relation syntaxique, catégorie sémantique de termes avec différents niveaux de granularité, etc.) ou de méthodes d'échantillonnage pour réduire le déséquilibre des données.

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

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.

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.