UNIVERSITÉ DE LORRAINE

Ecole doctorale IAEM Lorraine

Le résumé de la thèse

Découverte de connaissances considérant la littérature et les ontologies de domaine: Application aux maladies rares

Présentée et soutenue publiquement le 11 Juillet 2017

pour l'obtention du

Doctorat de l'Universit de Lorraine

(Specialite informatique)

par

Mohsen Hassan

Introduction

Une maladie rare, aussi appelée maladie orpheline, est, par définition, une maladie qui affecte un petit pourcentage de la population. Une maladie est considérée comme rare si elle affecte moins de 1000 ou 2000 personnes. Même si les maladies sont rares, il en existe une assez grande variété : on estime leur nombre entre 6000 et 8000. De plus, ces maladies sont souvent chroniques et très graves. Leur nombre cumulatif et leur gravité sont deux raisons qui rendent leur étude importante du point de vue de la santé. Cependant, plusieurs facteurs rendent cette étude difficile : (1) la difficultés à les diagnostiquer ; (2) le manque d'information les concernant dans des bases de données en ligne ; (3) le manque de connaissance scientifique les concernant. Une bonne connaissance des maladies rares pourrait réduire les mauvais diagnostiques et les retards dans le diagnostique et le soin. Pour cela, nous proposons d'extraire et de structurer les connaissances à partir de ressources textuelles disponibles.

Quelques bases de données comme Orphanet ou Orphadata décrivent les maladies rares. Malgré de très gros efforts, elles restent incomplètes et souvent en retard par rapport aux connaissances exprimées dans la littérature. En effet, il existe des millions de publications scientifiques portant sur ces maladies rares et le nombre de ces publications continue d'augmenter. Cela rend le processus d'extraction manuel difficile et très consommateur de temps. Une approche semi-automatique d'extraction d'information à partir de textes faciliterait cette tâche et représenterait l'information dans un format exploitable pour d'autres tâches, notamment la construction de connaissances.
Cette thèse vise à extraire l'information des textes et à utiliser ces informations pour enrichir des ontologies existantes dans le domaine médical, et plus particulièrement, des maladies rares. Nous apportons une contribution sur trois dimensions :

(1) l'extraction de relation à partir de textes entre maladies et phénotypes ;
(2) l'identification de phénotype complexes pour des maladies rares ;
(3) l'enrichissement d'une ontologie sur les maladies rares existante par les relations précédemment extraites

Une méthode Hybride pour l'extraction de relation entre maladies et phénotypes

Les relations Maladies-Phénotypes sont très importantes pour l'informatique biomédical puisqu'elles permettent une description fine d'une maladie er qu'elle peuvent aider au diagnostique dans le cas de soins cliniques. Dans ce contexte, nous proposons une méthode automatique, nommée SPARE*, pour extraire ces relations. SPARE* combine deux méthodes : une méthode originale basée sur les patrons syntaxiques nommée SPARE (Syntactic PAttern for Relationship Extraction), et une approche plus classique reposant sur un algorithme d'apprentissage, dans notre cas un SVM. Les résultats des deux approches sont combinés pour profiter de la précision élevée de SPARE et du bon rappel des SVM. Les deux approches

reposent sur des traits linguistiques, notamment syntaxiques, mais également des annotations sémantiques du domaine.

Pour apprendre et tester notre processus hybride d'extraction de relations entre maladies et phénotypes, nous avons construit un corpus annotés par ces entités. Ce corpus est constitué de 121 796 résumés d'articles scientifiques extraits de PubMed caractérisant 457 maladies rares distinctes. Les résumés ont été sélectionnés parce qu'ils contiennent le nom, ou un synonyme, d'une maladie rare selon la définition de l'ontologie Orphanet (ORDO : Orphanet Rare Disease Ontology).

Les 457 maladies rares ont été sélectionnées car elle répondent aux critères suivants : (1) elles sont associées dans la base OrphaData à des phénotypes (signes cliniques) ; (2) elles peuvent être mise en correspondance avec une maladie dans la base OMIM par le biais de l'UMLS ; (3) elles sont assosiées dans OMIM à un ensemble de phénotypes. Le corpus ainsi obtenu a une taille raisonnable et décrit des maladies qui ont des phénotypes à la fois dans OrphaData et dans OMIM. Les 121 796 résumés ont été découpés en 907 088 phrases. Chaque phrase a été annotée par Metamap pour repérer les maladies et les phénotypes. Ensuite, les phrases qui ne contiennent pas les deux types d'entités ont été retirées pour donner un corpus de 2 341 phrases. Finalement, chacune de ces phrases a été annotée manuellement par une relation (positive ou négative) entre les couples maladie-phénotype. 5 630 relations ont ainsi été annotées : 3 010 sont positives et 2 620 sont négatives.

SPARE repose sur l'identification de patrons syntaxiques construits à partir du chemin le plus court entre les deux entités maladie et phénotype dans le graphe de dépendances. SPARE se décompose en trois étapes. Tout d'abord, les patrons syntaxiques sont identifiés à partir du graphe de dépendances des différentes phrases comme étant les chemins les plus courts entre ces deux entités. Ces patrons sont ensuite généralisés. Dans ce processus de généralisation, les patrons sont regroupés sous un patron plus générique s'ils partagent les mêmes arcs selon le label et la direction. Ensuite, les patrons sont sélectionnés en fonction de leur support (le nombre d'occurrence de ce patron dans le corpus) et en fonction de leur capacité à prédire une relation (*ppv : positive predictive value*). Le *ppv* permet de sélectionner les patrons qui majoritairement introduiront des relations positives. Enfin, les patrons sélectionnés sont appliqués aux textes pour extraire de nouvelles relations maladie-phénotype.

La seconde méthode pour identifier les relations maladie-phénotypes est une méthode d'apprentissage supervisée, les SVM. Un SVM est utilisé pour classifier les occurrences maladie-phénotype en deux classes : les occurrence positives et les occurrences négatives. Les principales étapes pour mettre en place le SVM sont les suivantes : définir le corpus d'apprentissage, extraire les traits qui caractérisent le jeu de données, sélection des meilleurs traits à préserver pour la classification, et classer les nouvelles instances.

La combinaison de SPARE et de SVM se fait à partir des résultats de chacune des deux méthodes. L'idée est de tirer profit de la bonne précision de SPARE et du bon rappel du SVM. Différentes stratégies de combinaison des résultats ont été expérimentées. La solution retenue est celle qui donne les meilleurs résultats en

terme de F-mesure dont la valeur dans notre cadre expérimental vaut 0,81. C'est aussi la solution qui donne la meilleure valeur AUC-ROC de 0.79.

Reconnaissance de phénotypes complexes à partir des patrons syntaxiques

Les patrons syntaxiques précédemment identifiés par SPARE sont également utilisés pour découvrir de nouveaux phénotypes complexes en lien avec une maladie. Un tel phénotype n'est pas actuellement répertorié dans un dictionnaire. La méthode proposée consiste à sélectionner les patrons syntaxiques associés avec l'expression d'une relation maladie-phénotype et relaxer ces patrons du point de vue de la contrainte de phénotype pour pouvoir accepter des candidats qui n'ont pas été encore identifiés comme phénotype. Pour cela, une nouvelle mesure est associée aux patrons en plus du *ppv*, mesure que nous appelons la spécificité. La spécificité d'un patron mesure combien un patron est spécifique à l'expression d'une relation maladie-phénotype, lorsque l'on relaxe la contrainte de phénotype. On peut ainsi sélectionner et ne garder que les patrons les plus favorables à l'introduction d'une telle relation, ceux dont la valeur de spécificité est supérieure à un seuil donné. L'expérimentation de cette méthode sur les maladies rares est tout particulièrement intéressant pour permettre une caractérisation fine des maladies par leurs phénotypes.

La relaxation des patrons obtenus par SPARE permet donc d'identifier de nouveau phénotypes qui ne sont pas identifiés par les systèmes de reconnaissance d'entités nommées comme Metamap. Certains de ces candidats phénotypes peuvent être déjà présents dans une ontologie mais sous une forme différente et certains autres ne sont pas du tout répertoriés. Nous avons donc mis en place un processus de classification de ces candidats phénotypes en les comparant aux phénotypes présents dans HPO. Nous avons ainsi défini 5 catégories pour caractériser le niveau de correspondance entre le candidat et un phénotype répertorié : exacte, plus spécifique, plus générale, frère, aucune correspondance. Ainsi, si SPARE* identifie une phénotype déjà répertorié, la correspondance sera exacte, sinon, notre algorithme proposera un autre type de correspondance en comparant le candidat à tous les phénotypes répertoriés.

La stratégie de mise en correspondance repose sur un modèle sémantique vectoriel et compositionnel, complété par un ensemble de règles. Tout d'abord, un dictionnaire contenant tous les mots en lien avec la description des phénotypes est construit. Ensuite, les vecteurs sémantiques sont construits pour tous les mots impliqués dans la description des phénotypes. Pour les expressions, un vecteur résultat est construit à partir du sens individuel des mots. Enfin, la distance cosinus est utilisée pour trouver les phénotypes les plus proches dans HPO.

3 296 phénotypes parmi les 3821 phénotypes extraits par SPARE ont ainsi pu être reliés à certains des 11021 phénotypes de HPO. 525 phénotypes extraits n'ont pu être mis en correspondance, soit parce qu'ils sont nouveaux ou parce qu'ils sont trop bruités.

Spare a été utilisé pour enrichir Orphanet et Orphadata. Pour une maladie donnée, les phénotypes extraits par SPARE dans la littérature sont comparés aux phénotypes connus dans ces deux bases. Cette comparaison se fait en utilisant la même distance sémantique. Les phénotypes non déjà répertoriés sont alors proposés pour enrichir les connaissances sur la maladie.

Les structures de patrons pour la classification et l'enrichissement d'ontologie

Nous avons présenté une utilisation originale de l'exploration de textes et des structures de patrons pour fournir une nouvelle classification des maladies rares et enrichir une ontologie existante sur ces maladies. Cette classification est basée sur les descriptions phénotypiques des maladies rares. Elle utilise une connaissance représentée par une ontologie des maladies rares et une ontologie phénotypique.

Les structures de patrons classent les objets (par exemple, les maladies rares) en fonction de leurs descriptions (par exemple, leurs classes de maladie dans l'ontologie et les ensembles connus de phénotypes) dans une structure appelée treillis de concepts. Nous avons redéfini l'opérateur de structure de patrons pour prendre en compte à la fois une ontologie de maladies rares, c'est-à-dire une classification existante d'objets et une ontologie de phénotypes. Le treillis de concepts résultant est une classification des objets, qui regroupe les maladies rares qui partagent des phénotypes communs ou similaires selon l'ontologie des phénotypes. De ce fait, cette classification regroupe de nouveaux ensembles de maladies rares qui n'étaient pas proches dans l'ontologie de maladies rares initiale. Par conséquent, cette classification a suggéré de nouvelles classes pour enrichir l'ontologie initiale.
Nous avons expérimenté notre méthode sur la classification des maladies cardiaques rares qui contient 207 maladies rares. Étant donné que seules quelques maladies rares ont des phénotypes dans Orphadata, nous avons utilisé notre méthode SPARE*.

Le réseau résultant contient 4 829 concepts, un nombre de concepts beaucoup trop gros pour être pris en compte par les experts. Comme nous sommes intéressés par trouver des regroupements de deux maladies rares ou plus, nous avons d'abord écarté tous les concepts qui ne possédaient qu'une seule maladie dans leur extension. Il en résulte 4 662 concepts, ce qui fait encore un grand nombre de concepts. Par conséquent, nous avons fourni deux méthodes différentes de sélection de concepts pour trouver les plus intéressant. La première méthode utilise la *p-value* qui est une mesure statistique que nous avons utilisée pour mesurer la force de l'association entre deux maladies rares partageant des phénotypes. La seconde méthode, appelée *Gap*, est basée sur l'écart ou la dissemblance entre un concept et ses super-concepts.

Nous éliminons les concepts très similaires en termes de phénotypes à l'un de leurs super-concepts, car ce super-concept a un ensemble similaire de phénotypes mais regroupe plus de maladies rares dans son extension.
Plusieurs expériences ont été menées pour évaluer la méthode *Gap* et sa combinaison avec la méthode *p-value*. Cette évaluation est réalisée par la comparaison de ces méthodes avec une sélection aléatoire qui sert de méthode de

référence. La comparaison prend en compte la stabilité moyenne, la similarité avec l'ontologie initiale et les p-*valeurs* moyennes.