Apprentissage de patrons lexico-sémantiques d'acquisition de relations sémantiques

Thomas Rubiano
Stagiaire LIMSI CNRS
Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
Rue John von Neumann, Université Paris-Sud
91403 ORSAY, France

Résumé

Les méthodes d'extraction de relations sémantiques sont nombreuses mais elles nécessitent soit des corpus annotés soit une adaptation "manuelle" lorsqu'on veut identifier d'autres types de relations. Ces méthodes sont souvent spécialisées. Beaucoup utilisent l'apprentissage mais dans la plupart des cas on ne peut pas connaître les informations qui permettent d'extraire les relations visées. Certaines se basent sur l'acquisition incrémentale de patrons lexico-syntaxiques [Morin 1998]. Ce type d'approche se concentre sur le contexte dans lequel apparaît la relation. À partir d'exemples et en définissant ce contexte de manière symbolique, on peut induire des règles sur ce contexte. La *Programmation Logique Inductive* nous permet d'apprendre ces règles pouvant ensuite être traduites en patrons lexico-syntaxiques qui serviront à extraire de nouvelles relations. De plus, il est possible d'enrichir ces contextes d'informations sémantiques. Dans cette perspective, nous présentons les résultats d'une expérience préliminaire visant à apprendre des patrons *lexico-sémantiques* d'extraction de relations entre aliments et médicaments dans un corpus biomédical en anglais.

Abstract

Extraction methods of semantic relations need annotated corpus or manual adaptation when we want to identify other types of relations. These methods are often specialized. Many use machine learning but in most of cases we can't have access to the extraction informations. Some methods are based on incremental acquisition of lexical and syntactic patterns. This approach focuses on the context in which the relations appears. From example and by setting this context symbolically, rules can be induced. The Inductive Logic Programming (ILP) allows us to learn the rules which can be translated into lexical-syntactic patterns, and will be used to extract new relations. Moreover, it is possible to enrich these contexts with semantic informations. To this perspective, we present the results of an experiment to learn extraction lexical-semantic pattern of relations between food and drugs in a biomedical corpus in English.

Table des matières

1 Introduction **4**

2 Contexte applicatif et définitions **5**
2.1 Contexte applicatif 5
2.2 Définitions 6

3 État de l'art **7**
3.1 Introduction 7
3.2 Identification de relations 8
3.3 Acquisition de relations par apprentissage 10
3.4 Apprentissage Symbolique 10
3.5 Positionnement 12

4 Méthode proposée **12**
4.1 Programmation Logique Inductive 14
4.2 Langage d'hypothèse et background knowledge 15
4.3 Amorçage et définition des exemples 16
4.4 Traduction des règles en patrons lexico-sémantiques 17
4.5 Paramétrage d'Aleph 18

5 Matériel **19**
5.1 Ressources 19
5.2 Corpus 19
5.3 Pré-traitement à l'aide de la plateforme Ogmios 20

6 Expérimentations et Résultats **21**
6.1 Expérimentations 21
6.2 Résultats 21

7 Conclusions et Perspectives **24**
7.1 Conclusions 24
7.2 Perspectives 24

1 Introduction

Les textes de spécialité contiennent des informations sur les interactions entre différents concepts. Cela nécessite d'extraire des relations sémantiques et pour cela on dispose de différentes approches qui peuvent s'appuyer sur diverses méthodes : des patrons lexico-syntaxiques, de l'apprentissage supervisé, etc. Les patrons lexico-syntaxiques décrivent les contextes lexicaux et syntaxiques caractérisitques des relations qu'entretiennent les termes. Ces relations sont ainsi extraites avec une bonne précision mais les patrons nécessitent d'être définis en général manuellement. Dans notre travail nous proposons d'utiliser une méthode d'apprentissage supervisé pour définir automatiquement les patrons d'une relation en s'appuyant sur des exemples issus de ressources terminologiques ou fournis par des patrons d'amorçage. Pour cela, nous mettons en oeuvre une méthode basée sur la Programmation Logique Inductive. En effet, la Programmation Logique Inductive, grâce à son aspect symbolique, semble adaptée pour cette tâche : les règles générées peuvent (si la structure du contexte est correctement définie) être interprétées comme des patrons d'acquisition de relations sémantiques. Nous testons notre approche en l'appliquant à l'apprentissage de patrons qui permetteront d'identifier les relations aliment/médicament au sein d'un corpus de résumés d'articles scientifiques en anglais. Après avoir introduit le contexte applicatif et les notions importantes dans notre travail, nous présentons les différents travaux effectués dans le domaine d'acquisition de relations pour ensuite nous concentrer sur l'apprentissage symbolique. Puis, nous présenterons la méthode mise en oeuvre et les expérimentations effectuées. Après quelques analyses, nous ferons le point sur ce qui a été fait et les perspectives envisagées.

2 Contexte applicatif et définitions

2.1 Contexte applicatif

La qualité de vie et la santé des personnes dépendent de plusieurs facteurs, parmi lesquels se trouvent par exemple l'environnement, le contexte social, psychologique et économique, la médecine et la qualité des soins médicaux, la situation géographique et l'alimentation. Cela crée une situation d'autant plus complexe que ces facteurs interagissent naturellement entre eux.
Ainsi, le contexte applicatif de notre travail concerne les interactions qui peuvent exister entre les pathologies, les médicaments et l'alimentation (aliments et boissons). Si le lien entre les médicaments et les pathologies a déjà fait l'objet de multiples travaux, leur relations avec l'alimentation a été beaucoup moins étudiées, et surtout très peu recensées dans les bases de connaissances biomédicales.

Pour appuyer ce propos prenons l'exemple de l'enzyme CYP3A4. En 1998, plusieurs chercheurs ont montré que le jus de pamplemousse, et le pamplemousse en général, est un puissant inhibiteur du CYP3A4. De ce fait, la consommation de pamplemousse pendant un traitement médicamenteux peut diminuer l'élimination par l'organisme du médicament, et en augmenter la biodisponibilité. Cela peut provoquer un surdosage qui peut être fatal. La première publication à ce sujet date de 1991[Bailey et al. 1991], ce qui était la première fois que l'on observait cliniquement une interaction entre la nourriture et les médicaments. Malgré cela, il a fallu attendre plusieurs cas de décès pour que cet effet soit vraiment pris en compte. De plus, les seules informations concernant les liens entre l'alimentation et les médicaments enregistrées dans la base DrugBank le sont sous forme de phrase, sans formalisation.

2.2 Définitions

Relation : Dans ce rapport, nous nous intéressons aux relations binaires (nous verrons plus tard pourquoi) et plus particulièrement aux relations aliments/médicaments. Par exemple dans la phrase :

grapefruit juice interacts with a variety of medications

Les termes *grapefruit juice* et *medications* entrent en relation.

Patron lexico-sémantique : Un patron lexico-sémantique décrit le contexte lexical et sémantique dans lequel des termes s'insèrent et entrent en relation. Certains patrons ont une fréquence d'apparition plus élevée en fonction du domaine. Lorsque le patron est pertinent il permet d'extraire des relations à moindre coût [Hearst 1992].

This study evaluated the effect of grapefruit juice on the pharmacokinetics of nilotinib in 21 healthy male participants.

lexical/sémantique :	effect	of	FOOD	on	DRUG
morpho-syntaxique :	NN	IN	catSem	IN	catSem

FIGURE 1 – Exemple de patron lexico-sémantique

Patron partiel : Lorsque nous parlerons de patrons *partiels*[Saeger et al. 2009] il s'agira de patrons décrivant le contexte autour d'un terme de la relation. Dans l'exemple précédent, un patron partiel pourrait être :

lexical/sémantique :	effect	of	FOOD
morpho-syntaxique :	NN	IN	catSem

FIGURE 2 – Exemple de patron partiel

Ici, le patron partiel décrit le contexte dans lequel l'aliment apparaît.

3 État de l'art

En introduction, nous définirons plus précisément une relation. Nous prendrons connaissance des différents travaux effectués pour identifier ces relations. Puis nous nous intéresserons plus particulièrement à l'acquisition de ces relations par apprentissage pour ensuite nous centrer sur l'apprentissage symbolique. Enfin nous présenterons notre positionnement.

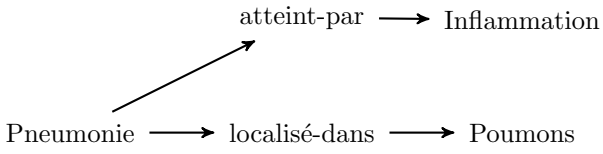
3.1 Introduction

Une relation peut relier des entités nommées, comme des noms de personnes, de lieux, des dates, ou des termes, comme des noms de gènes, d'aliments, de médicaments. Une relation peut être restreinte à des types sémantiques particuliers (ex. la relation *auteur_de* n'intervient qu'entre un nom de personne et un titre) ou au contraire, elle peut exister pour tous types (ex. la relation d'hyperonymie *est_un*).

Parmis les types de relations sémantiques les plus utiles dans les textes de spécialité[VN 2013] :

- les relations *taxinomiques* : Ce type de relation structure des termes dans une arborescence.
  - Les relations d'*hypéronymie* (*est_un*) relient un terme général à un terme spécifique.
  - Les relations partitives (*méronymie* ou partie-tout) sont utilisées pour définir une relation entre deux termes ou l'un est une partie de l'autre. (ex. *bras* est un méronyme de *corps*)
- Les relations sémantiques lexicales regroupent deux types de relations :
  - les relations de *synonymie* ou d'équivalence qui relient les termes possédant le même sens [Hamon 1998, Jacquemin et al. 1997].
  - les relations d'*antonymie* ou d'opposition qui relient les termes ayant des sens contraires.

- Les relations inter-hiérarchiques (*transversales*) relient les termes appartenant à des branches différentes de l'arborescence hiérarchique. Elles sont dépendantes des domaines, des corpus et même des applications [Grabar et al. 2004]. Elles fournissent des informations précieuses et spécifiques au domaine. (ex. une *pneumonie* est localisée dans les *poumons* qui subit une *inflammation*)



3.2 Identification de relations

L'identification de relations d'une phrase peut être réalisée à l'aide d'informations syntagmatiques et/ou paradigmatiques [VN 2013].

Les informations syntagmatiques peuvent être obtenues en étudiant les mots qui partagent le même contexte (qui co-ocurrent), on parle d'affinités du premier ordre selon [Grefenstette 1994]. Différentes affinités peuvent être extraites dépendamment de ce qui est défini comme "contexte" :

- [Church et Hanks 1990] ont montré qu'en calculant l'information mutuelle (*the mutual information*[1]) entre des mots contenus dans une fenêtre de taille fixe dans un large corpus, il est possible de reconnaître des affinités entre des paires telles que *docteur...nurse* et *save...from*. D'autres approches similaires ont été utilisées par la suite (ex. *Likelihood Ratio* [Dunning 1993]).

- [Smadja 1993] présente une technique pour reconnaitre des expressions, patrons et *collocations* en examinant les catégories grammaticales des mots dans une fenêtre donnée. La méthode proposée construit un tableau répertoriant tous les mots apparaissant autour du mot cible en respectant leurs positions. Les mots apparaissant souvent dans une position fixe sont considérés comme des co-occurrents possibles. Par exemple *hostile...takeover* est souvent apparu comme une *collocation* avec un positionnement fixe entre les deux mots, tandis que *federal...takeover*, quoiqu'apparaissant souvent dans la même fenêtre de texte, n'ont pas été retenus par les filtres de positionnement et ne sont donc pas considérés comme co-occurants.

- [ACL 1993] détermine des sous-catégorisations à partir de texte à l'aide d'évaluation statistiques des mots qui apparaissent autour d'un verbe donné.

- L'utilisation de patrons lexico-syntaxiques [Hearst 1992] est une méthode efficace d'acquisition des relations au sein d'une phrase, elle obtient une bonne précision mais un rappel moyen. C'est une approche robuste mais coûteuse car manuelle. Elle a été mise en oeuvre pour acquérir des relations d'hyperonymie [Hearst 1992, Morin 1999], des relations de synonymie [Weissenbacher 2004, Mccrae et Collier 2008], mais aussi des relations spécifiques à un domaine comme les facteurs de risque d'une maladie [Hamon et al. 2010a]. L'expérience montre que ces patrons sont plus ou moins pertinents et doivent toujours être adaptés.

Les informations paradigmatiques (affinités du second ordre) sont acquises

1. La formule pour calculer l'information mutuelle est I(x y) = log(P(x y)/(P(x) P(y))) où P(x y) est la probabilité conjointe des évènements x et y et P(x) et P(y) sont les probabilités de chaque évènement. La valeur croît lorsque x et y co-occurent et sont rares

en examinant les mots qui apparaissent dans les mêmes contextes délimités par une taille donnée (on parle de fenêtre en nombre de mots). Les mots partageant une affinité du second ordre doivent rarement apparaître ensemble, mais leur environnement est similaire. Cette approche est très intéressante pour l'acquisition de relation de synonymie. [Grabar et Zweigenbaum 1999] obtiennent des collocations "*gene expression*" et "*genic expression*" où "*gene*" et "*genic*" sont de la même famille morpho-sémantique. Les autres s'intéressent ainsi aux paires de mots morphologiquement reliés en étudiant les inclusions lexicales et proximités sémantiques entre termes [Grabar et Zweigenbaum 2003, Bodenreider et al. 2001, IS 2005]. Ils considèrent comme morphologiquement liés, une paire de mots qui sont dérivés de la même racine (ex "*symbio*se" / "*symbio*tique"). Pour les déterminer, un moyen simple est d'examiner leur préfixe commun le plus long. Cependant une telle approche peut conduire à beaucoup de bruits : par exemple "*admi*nistratif" / "*admi*ssion" ont une chaîne initiale commune mais ne sont pas morphologiquement reliés.

[Harris 1971] a présenté sa conception distributionnaliste de la sémantique reprise ensuite par Sager et [Habert et Nazarenko 1996]. Ces méthodes d'évaluation relèvent d'une approche quantitative des données qui ne laisse que peu de place à l'interprétation des rapprochements générés. En effet, l'analyse distributionnelle s'appuie sur un principe de corrélation entre les contextes dans lesquels les mots apparaissent (leur *distribution*) et leur contenu sémantique. L'observation des contextes dans lesquels apparaissent les mots d'un corpus permet d'établir des *classes sémantico-distributionelles*. Les travaux de [MH 2013] visent à mieux comprendre les fonctionnements qui régissent les rapprochements distributionnels en s'intéressant aux relations de synonymie, antonymie, hyperonymie et méronymie de manière qualitative. Pour [MH 2013], "les bases distributionnelles sont le résultat d'une mise en oeuvre à grande échelle du principe de substituabilité".

Donc, les méthodes fondées sur les informations paradigmatiques ne sont pas appropriées pour extraire et catégoriser des relations dans un domaine spécifique [Minard 2012] (ex. relations transversales). Par exemple, si on veut extraire les relations entre un médicament et une maladie, le fait que les deux entités co-occurrent un certain nombre de fois dans le corpus, ne permettra pas de classer la relation correctement.

3.3 Acquisition de relations par apprentissage

Les différentes méthodes d'acquisition de relations utilisant l'apprentissage automatique peuvent être vues comme des problèmes de *classification* car elles produisent des méthodes permettant de *classer* les divers couples de mots observés comme décrivant ou non la relation cible. On peut regrouper ces méthodes de classification en différentes familles selon les attributs utilisés pour les éléments de la relation recherchée : on distingue donc les travaux se basant sur l'aspect fréquentiel des mots, ceux qui exploitent des indices structurels, symboliques et ceux dits hybrides qui utilisent les deux approches [Claveau et Sébillot 2004]. Dans notre travail, nous nous sommes appuyé sur une approche par apprentissage symbolique, bien qu'en pratique, plusieurs méthodes d'acquisition existantes se fondent à la fois sur les deux aspects, numérique et symbolique. Les études de type "numériques" visent à extraire des informations syntagmatiques et paradigmatiques comme vu précédemment 3.2. Reposant sur des informations fréquentielles, l'extraction de termes s'insérant dans la relation recherchée est réalisée au niveau du corpus pris dans son ensemble. Des indices statistiques d'association sont un outil couramment utilisé pour mettre au jour les relations syntagmatiques en pointant les mots qui cooccurrent dans une zone de texte de manière statistiquement significative.

Les techniques d'apprentissages supervisés utilisées sont les SVM (Machines à Vecteurs de Support) [Uzuner et al. 2010, Roberts et al. 2008]. D'autres systèmes utilisent des classifieurs basés sur les CRF [Lafferty et al. 2001, Sahay et al. 2008] ou sur des réseaux de neurones [Rosario et Hearst 2004]. Les attributs utilisés pour représenter les relations sous forme vectorielle peuvent être des attributs de surface, des attributs lexicaux, des attributs morpho-syntaxiques et syntaxiques : les attributs lexicaux (par exemple les mots de la phrase) et de surface (par exemple la distance entre deux entités) ne sont pas toujours suffisants pour identifier correctement une relation [Minard 2012]. L'information syntaxique ou sémantique peut améliorer la précision du système. Les informations sémantiques transposées sous forme d'attributs peuvent provenir de ressources pour la langue générale ou de ressources d'un domaine de spécialité.

3.4 Apprentissage Symbolique

Bien que dans la plupart des cas, les patrons sont définis manuellement, ils peuvent aussi être appris à l'aide d'un corpus annoté. Les méthodes qui reposent sur la définition manuelle de patrons sont généralement efficaces uniquement en précision. Les patrons construits ou extraits à partir de phrases dans lesquelles les entités en relation sont très éloignées, sont trop spécifiques. Il est nécessaire de définir des patrons qui ne soient ni trop spécifiques, ni trop génériques. Par exemple, un patron entièrement lexicalisé aura tendance à être trop spécifique, alors qu'un patron formé uniquement des catégories

morpho-syntaxiques des mots sera au contraire trop générique. Il est important que, lors de l'apprentissage des patrons, le contexte soit étudié à la fois d'un point de vue morphologique, syntaxique, lexical et sémantique [Minard 2012].

Les patrons lexico-syntaxiques sont définis à partir d'observations en corpus. L'utilisation d'une méthode d'affinement des patrons lexico-syntaxiques par apprentissage automatique permet de spécialiser les observations réalisées et d'obtenir de bons résultats [Morin et Martienne 1999]. Des relations transversales peuvent également être acquises grâce à cette stratégie [Røst et al. 2010].

Les patrons lexico-syntaxiques peuvent être aussi construits par apprentissage supervisé. Les indices trouvés en corpus dans le contexte sont exploités pour inférer des patrons par *programmation logique inductive* [Martienne et Morin 1999, Claveau et Sébillot 2004]. Des relations d'hypéronymie, dans le premier travail, ou des relations entre des verbes et des noms, dans le second, sont ainsi identifiées.

Des contextes plus larges peuvent être exploités pour acquérir des relations entre termes. Ainsi, des contextes riches en connaissances (Knowledge-Rich Context[Schumann 2011]) sont définis selon comme étant un contexte qui contient des termes d'un domaine spécialisé et des modèles (patterns) de connaissances [Meyer 2001]. Des relations sémantiques peuvent alors être acquises en analysant ces contextes.

[Claveau 2003] a proposé et évalué une méthode d'extraction de couples N-V basée sur une méthode d'apprentissage symbolique supervisé utilisant la programmation logique inductive (PLI) [Muggleton et Raedt 1994]. Dans cette méthode, la PLI sert à générer un *classifieur* (un ensemble de règles Prolog) de manière supervisée, c'est-à-dire à l'aide d'exemples en contexte et de contre-exemples. Les règles obtenues sont ensuite utilisées comme patrons d'extraction. Les résultats de cette approche sont de bonne qualité. Cependant, ces patrons étant a priori propres à chaque corpus, il est nécessaire de reconduire la phase d'apprentissage par PLI pour tout nouveau texte. Comme toutes les approches supervisées, cette première présente l'inconvénient de nécessiter la construction d'un jeu d'exemples et de contre-exemples propres au corpus à traiter. Cette phase, essentiellement manuelle (et exigeant l'aide d'un expert dans certains cas), est très coûteuse, et conduit cette technique d'acquisition à ne pas répondre entièrement au souci de portabilité et d'automaticité.

Pour répondre à nos problèmes [Claveau et Sébillot 2004] proposent deux variantes de cette méthode d'acquisition de relations qui remédient à ce problème et remplissent ainsi les différentes exigences : bonne qualité des résultats, interprétabilité linguistique et automaticité du processus. Pour ce faire, en conjonction de cette approche symbolique supervisée, ils utilisent une technique reposant sur une approche différente de l'extraction : l'approche statistique.

Les systèmes d'extraction hybrides (statistique et symbolique) résultants sont entièrement automatiques et ne nécessitent plus de fournir manuellement des exemples. Ces deux variantes non supervisées obtiennent des résultats similaires à la méthode originale.

Parmis les méthodes hybrides, on trouve les *Markov Logic Network* (MLN) [Richardson et Domingos 2006]. Elles acquièrent préalablement des données statistiques qui vont leur permetttre de pondérer chaque clauses ou contraintes données en *background knowledge* (voir la section 4.2).

3.5 Positionnement

[Morin et Martienne 1999] utilisaient le Programmation Logique Inductive pour affiner des patrons lorsque [Morin 1999] proposait une méthode d'acquisition ad hoc pour identifier des patrons lexico-syntaxiques. Ces derniers étaient généralement définis manuellement.

La volonté d'apprendre directement des patrons lexico-sémantiques nous a conduit naturellement à nous positionner dans le cadre des approches structurelles et plus précisément celui de l'apprentissage symbolique.

Comme le fait remarquer [Claveau et Sébillot 2004], les méthodes d'acquisitions à partir d'indices numériques (même si elles sont portables et automatiques) souffrent d'un manque d'interprétabilité. Il est souvent difficile de comprendre pourquoi un couple d'éléments cooccurrants a été retenu et pas un autre, le seul indice fourni à ce sujet étant généralement un score statistique. Les approches distributionnelles proposent des regroupements de mots sans toutefois permettre d'identifier le type précis des relations existantes entre ceux-ci.

Notre positionnement est réalisé en pratique par l'utilisation d'une méthode d'apprentissage supervisée : nous utilisons la programmation logique inductive pour produire règles décrivant des patrons qui permettent d'acquérir des relations. Nous présentons dans la section 4, la méthode que nous proposons.

4 Méthode proposée

Nous pouvons distinguer les différentes briques qui composent notre méthode : l'élément central est un processus d'apprentissage par Programmation Logique Inductive (*PLI* voir section 4.1). Celui-ci requiert des données d'apprentissage (Exemples positifs (E+), négatifs (E-) (voir section 4.3) et des connaissances initiales *"Background knowledge"* voir section 4.2). Nous avons donc besoin de relations en contexte dans notre corpus pour acquérir ces exemples. Pour ce faire, nous disposons de différentes possibilités d'amorçage (voir

section 4.3) :

- acquisition manuelle de relation directement à partir du corpus
- acquisition en contexte à l'aide de patrons lexico-sémantiques connus (1 - en vert dans la figure 3)
- projection de relations connues sur le corpus (à l'aide d'une ressource terminologique par exemple) (2 - en bleu dans la figure 3)

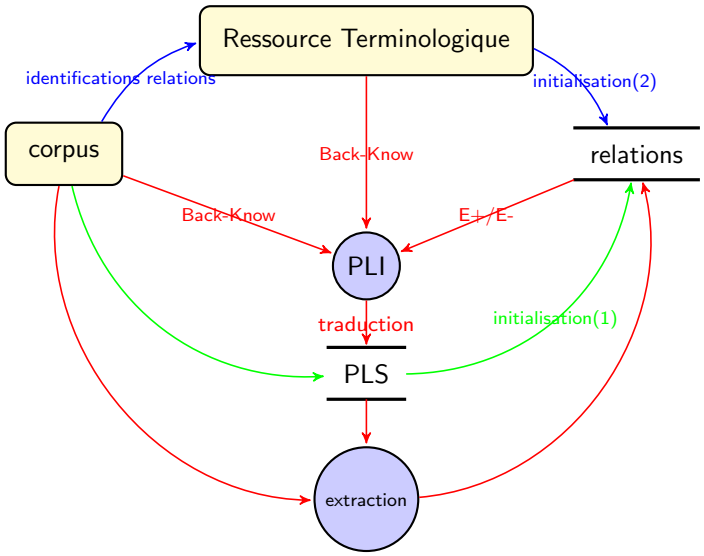


FIGURE 3 – Méthode proposée

Une fois l'apprentissage amorcé, il est possible de réitérer ce processus en utilisant les relations extraites à partir des patrons appris lors de l'occurrence précédente. Ce mécanisme devrait, en fonction de la méthode d'initialisation choisie, nous permettre d'obtenir soit de nouveaux patrons décrivant de nouveaux contextes (et potentiellement de nouvelles relations), soit des patrons plus précis décrivant le même contexte (qui n'extrairont pas forcément de nouvelles relations). Nous verrons dans la section 4.4 comment nous traduisons nos règles en patrons.

Les premières expériences que nous avons réalisées montrent qu'il semble préférable de choisir un paramétrage ainsi qu'une initialisation en fonction de l'objectif final : Si le but est d'obtenir de nouveaux contextes propices à l'apparition de relations cibles, il est nécessaire d'acquérir les premiers exemples manuellement ou à l'aide de relations connues (il serait également envisageable d'autoriser un peu de bruit afin d'augment le rappel). Si, au contraire, le but est de valider ou d'affiner un modèle connu, on peut utiliser ce modèle lors de l'initialisation (il serait alors nécessaire de diminuer le bruit autorisé, la section 4.3 discute ce point).

4.1 Programmation Logique Inductive

La PLI [Muggleton et Raedt 1994] se situe à la croisée de la programmation logique et de l'apprentissage automatique. Elle est utilisée pour induire des règles générales (sous forme de clauses de Horn) expliquant un concept à partir d'exemples, de contre-exemples de ce concept et d'un ensemble d'informations préexistantes appelé *background knowledge*.

L'avantage majeur de la PLI est de permettre l'apprentissage à partir d'exemples relationnels, usuellement exprimés en Prolog. C'est cette expressivité, à la fois en entrée et en sortie du processus d'apprentissage, qui rend la PLI intéressante pour traiter certains problèmes difficilement exprimables.

Un *langage d'hypothèses* est également donné à l'algorithme de PLI. Il est utilisé pour définir la forme attendue des règles générées. Ce langage assure ainsi de n'obtenir que des règles bien formées et pertinentes au regard de la tâche d'apprentissage visée. En fonction de ce langage, l'objectif de la PLI est donc d'induire des règles qui couvrent (c'est-à-dire expliquent à l'aide de règles) un maximum d'exemples et aucun contre-exemple (ou très peu selon le bruit autorisé). Lors de ce parcours, le choix d'une hypothèse est fait selon une fonction de score, qui dépend le plus souvent du nombre d'exemples positifs et négatifs couverts par l'hypothèse.

Le logiciel de PLI utilisé lors de nos expérimentations est *ALEPH*, une implémentation en Prolog réalisée par Ashwin Srinivasan. Son fonctionnement peut être simplement décrit par l'algorithme donné ci-dessous :

Data: $E+$ (exemples positifs pour apprentissage)
Result: H (ensemble de clauses maximisant la couverture des exemples)
while $E+$ *n'est pas vide* **do**
    choisir aléatoirement un exemple positif $e+$ dans $E+$;
    définir un espace de recherche d'hypothèses Eh à partir de $e+$ et du langage d'hypothèses;
    parcourir l'espace de recherche Eh à la recherche de la clause h maximisant une fonction de score;
    ajouter h à l'ensemble H et ôter de $E+$ les exemples couverts par h;
end

Algorithm 1: Algorithme d'*ALEPH*

Cette méthode d'apprentissage a été choisie pour deux raisons liées à son pouvoir expressif : l'encodage des relations syntaxiques, lexicales et sémantiques se fait très naturellement avec des prédicats (voir 6.2). Par ailleurs, le classifieur obtenu, c'est-à-dire l'ensemble de règles H, est facilement interprétable et se prête donc aisément à une traduction en patrons lexico-sémantiques (voir définition à la section 2.2).

4.2 Langage d'hypothèse et background knowledge

On définit le *langage d'hypothèse* dans le *background knowledge* en vu d'un apprentissage symbolique. Le *langage d'hypothèse* se compose de :

- prédicats décrivant un mot :
  - *has_wordform(+wordId,#wordform)* où *wordId* est l'identifiant du mot et *wordform* est la forme fléchie du mot.
  - *has_lemma(+wordId,#lemma)* où *lemma* est la forme non fléchie du mot.
  - *has_postag(+wordId,#postag)* où *postag* est une étiquette morpho-syntaxique.
  - *has_term(+wordId,#catSem)* où *catSem* est un type sémantique (dans notre cas nous pouvons nous concentrer sur deux types sémantiques (*food* et *drug*). On verra plus tard qu'il est possible d'enrichir l'ensemble des types sémantiques afin d'obtenir des patrons plus précis (ex. effets indésirables, sous-groupes d'aliments etc...).
- prédicats décrivant la position de chaque mot dans une phrase :
  - *pred(B,A)* veut dire que le prédecesseur de A est B.
  - *sentence_beginning(A)* pour initialiser le premier mot d'une phrase.

Ce langage nous permet de décrire chaque phrase du corpus dans notre *background knowledge* comme présenté à la figure 4.

```
has_wordform(s3w26,adverse).   has_postag(s3w26,jj).      has_lemma(s3w26,adverse).
pred(s3w27,s3w26).             has_wordform(s3w27,effects). has_postag(s3w27,nns).
has_lemma(s3w27,effect).       pred(s3w28,s3w27).         has_wordform(s3w28,due).
has_postag(s3w28,jj).          has_lemma(s3w28,due).      pred(s3w29,s3w28).
has_wordform(s3w29,to).        has_postag(s3w29,to).      has_lemma(s3w29,to).
pred(s3w30,s3w29).   has_wordform(s3w30,druginteractions). has_postag(s3w30,catSem).
has_lemma(s3w30,druginteraction).
```

FIGURE 4 – Portion du *background knowledge*

Il est également possible d'ajouter certaines contraintes ou d'autres définitions utiles pour l'apprentissage. Par exemple si nous voulons définir une distance entre les mots :

```
distance(A,B,1) :- pred(B,A).
distance(A,C,AC) :- distance(A,B,AB),
        distance(B,C,BC),
        AC is AB + BC,
        AC ≤ 4, !.
```

FIGURE 5 – Définition et contrainte sur la distance entre les mots

Une notion de distance entre les mots décrit à la figure 5 a été introduite. Cependant nous avons constaté que le processus nécessitait un temps de calcul beaucoup trop conséquent et finalement nous avons abandonné l'utilisation de cette notion. En effet, la programmation logique semble difficilement adaptable pour effectuer des manipulations d'expressions numériques.

4.3 Amorçage et définition des exemples

Comme dit précédemment, la méthode proposée devra induire des patrons d'extraction de relation à partir d'exemples et de connaissances de bases (*background knowledge*). Aussi, pour acquérir ces exemples positifs et négatifs en contexte, plusieurs possibilités s'offrent à nous :

1. distinguer différentes classes de relations, il s'agit d'apprentissage *multiclass*. Il n'y a pas d'exemples négatifs, tout exemple appartenant à une classe est positif pour cette dernière mais négatif pour les autres. Par exemple nous pouvons établir 3 classes de relations : relation globale, augmentation de l'effet du médicament et diminution de l'effet du médicament. Cette méthode nécessite plus de vérifications pour classer les relations en fonction des patrons utilisés.
2. distinguer la relation visée de tout le reste (*One Vs All*). Ainsi, pour l'amorçage, chaque couple aliment/médicament apparaissant au sein d'une même phrase et déjà présent dans notre ressource terminologique, est une relation acquise donc présente dans un contexte propice. Cette technique nécessite des relations connues pour amorcer l'apprentissage des patrons et elle peut faire diminuer la précision.

Cependant, lors de nos tests, nous avons constaté que, malgré un paramétrage ad hoc, l'outil d'apprentissage symbolique ne générait pas le même résultat pour un même jeu de donnée en entrée. Une analyse en profondeur de l'outil de PLI a montré que l'approche *multiclass* nécessite que les exemples positifs soient pris dans un ordre aléatoire lors de l'apprentissage [2].

Afin de ne pas trop dégrader la précision, nous avons choisi d'utiliser des patrons connus et de vérifier si les relations extraites sont présentes dans notre ressource terminologique. Cette méthode permet d'accroitre considérablement la précision mais diminue fortement le rappel. Elle sera plus utilisée dans le cadre de l'affinement de patron.

2. *http* : *//www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html#SEC*36

4.4 Traduction des règles en patrons lexico-sémantiques

La fin d'une itération est marquée par la traduction des règles apprises par induction en patrons lexico-sémantiques. Deux questions peuvent se poser lors de la traduction :

1. est-ce que les règles peuvent être traduites et exploitées en patrons partiels ? si oui, comment ? (voir section 2.2 pour la définition d'un patron partiel)
2. devons-nous être aussi stricte que les règles générées (surtout vis-à-vis des distances entre les mots comme mentionné en section 6.2) ?

1. Cela dépend de la règle. Certaines règles peuvent être composées de deux patrons partiels dont voici un exemple :
   Dans ce cas il est possible de générer les deux formes mais cela peut

rel(A,B,relation) :-
has_lemma(C,of) ∧ pred(B,C) ∧ pred(A,D).
peut générer un patron de la forme :

C	←	B	D	←	A
of		[catSem]	...		[catSem]

ou de la forme :

D	←	A	C	←	B
...		[catSem]	of		[catSem]

FIGURE 6 – Exemple de patrons partiels

apporter du bruit. Une autre méthode envisageable est d'observer les couvertures de chacun des patrons partiels et de choisir le cas où l'intersection de couverture d'exemple est la plus élevée. Enfin, si la règle est partielle comme lors de l'exemple en figure 2, nous pouvons toujours observer l'intersection de couverture d'exemples avec les autres règles partielles pour construire une règle complète.

2. Nos expériences ont montré que les règles générées étaient souvent trop strictes du point de vue des distances. Nous avons donc remédié à ce problème en acceptant que les mots décrits dans notre règle soient plus ou moins éloignés les uns des autres. Par exemple, nous pouvons accepter un déterminant avant le terme, ou un adjectif. Pour l'exemple précédent, D est un mot qui peut être n'importe quoi, il n'y a aucune contrainte lexicale, syntaxique ou sémantique. Nous considérons que cet espace entre les deux termes B et A peut être plus large et nous acceptons un maximum de 4 mots dans nos tests. De plus, contrairement à notre langage d'hypothèse, les patrons prennent en compte les listes de termes.

4.5 Paramétrage d'Aleph

Le paramétrage de l'outil d'apprentissage logique est également très important. Il est nécessairement à établir suivant les besoins. En effet, si la fonction de score est trop stricte ou l'espace de recherche d'hypothèse trop petit, il se peut qu'aucune clause ne soit selectionnée. Alors l'apprentissage ne pourra pas amorcer une autre itération. Il est donc impératif d'ajuster les paramètres. Nous avons identifié quelques paramètres importants :

- alors qu'il est indispensable d'obtenir le même résultat pour la même configuration et les mêmes fichiers en entrée, par défaut l'algorithme d'apprentissage pioche au hasard les exemples à couvrir parmis ceux donnés (voir algorithme 1). Nous avons donc forcer l'ordre d'apprentissage des exemples.
- dans le même contexte il est préférable d'appliquer l'algorithme sur tous les exemples et non un sous-ensemble (aléatoire lui aussi).
- la taille de l'espace de recherche d'hypothèse (*bottom clause*) est importante dans notre cas car nous recherchons des contextes plus ou moins large. Cet espace de recherche correspond au nombre de contraintes décrites par des clauses autour d'une relation. Plus la taille est grande plus l'algorithme regardera la phrase dans son ensemble.
- les clauses sont selectionnées à l'aide d'une fonction de score. Ce score peut être calculé de différentes manières décrites dans le manuel d'Aleph [3]. Les principaux paramètres de cette fonction sont le nombre d'exemples positifs et négatifs couverts. Cependant, nous avons vu qu'il est difficile de définir des exemples négatifs pertinents. Il se peut donc que cette fonction influe négativement nos résultats. Une solution est d'accepter une certaine valeur de score (particulièrement bas) et d'établir une borne inférieur (N) assez haute en ne selectionnant que les clauses dont le nombre de couverture d'exemple positifs est supérieur à N.

3. *http* : *//www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html#SEC*22

5 Matériel

Nous décrivons dans cette section les ressources et le corpus utilisé lors de nos expériences.

5.1 Ressources

Les ressources apportent des informations essentielles pour notre apprentissage. Certaines vont nous permettre d'annoter notre corpus et d'augmenter la précision de nos patrons. Il est nécessaire de choisir (ou constituer soi-même) des ressources de qualité car la moindre erreur peut avoir des conséquences négatives sur les résultats.

Medline est une base de données bibliographique regroupant la littérature relative aux sciences biologiques et biomédicales. La base de données contient plus de 24 millions d'articles en anglais, référencés provenant de plus de 5000 sources différentes (revues en biologie et en médecine) dont les plus anciennes remontent à 1902. Notre corpus est constitué de certains de ces articles.

RxNorm [RxNorm 2009] est une ressource terminologique répertoriant tous les médicaments vendus aux États-Unis, elle fait partie des terminologies *UMLS*[4][National Library of Medicine 2003] et est maintenue par le *NLM*[5]. Cette ressource a préalablement été adaptée pour annoter des textes cliniques dans le cadre du *Challenge I2B2* [Hamon et al. 2010b]. Elle va essentiellement nous servir à identifier les médicments cités dans le texte.

DrugBank [Wishart et al. 2006] est une base de données publique et gratuite, accessible en ligne, concernant la bio-informatique et la chémoinformatique. *DrugBank* contient en 2013 environ 6 800 entrées de médicaments et près de 600 intéractions aliments/médicaments détectées. Elle va nous permettre d'identifier les termes qui nous intéressent : aliments et médicaments. Les intéractions répertoriées vont nous servir d'amorçage comme vu en section 4.3.

Aliments pour les ressources alimentaires nous avons constitué une liste contenant 542 termes d'un site producteur de contenu éducatif[6] ainsi qu'une ressource USDA[7].

5.2 Corpus

L'ensemble de résumés d'articles composant le corpus est récupéré à l'aide de la requête suivante sur *Medline* :

4. Unified Medical Language System
5. United States National Library of Medicine
6. http ://www.enchantedlearning.com/wordlist/food.shtml
7. http ://ndb.nal.usda.gov/ndb/search/list

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

Cette requête récupère 642 résumés (soit 137 000 mots) ce qui cible plus particulièrement les relations aliments/médicaments. Elle constitue un premier filtre afin de diminuer le bruit. Nous verrons plus tard que ce filtrage a un impact lors de l'apprentissage des patrons lexico-sémantiques.

5.3 Pré-traitement à l'aide de la plateforme Ogmios

Afin d'obtenir les informations morpho-syntaxiques et sémantiques nécessaires à l'apprentissage des patrons, un pré-traitement est réalisé. Le corpus est préalablement segmenté en phrases puis en mots. Chaque mot est lemmatisé et étiqueté de la façon suivante :

1. une lemmatisation consiste à faire passer le mot de la forme fléchie à la forme non fléchie (en vert dans la figure 7).
2. un premier étiquetage des catégories grammaticales ou *étiquetage morpho-syntaxique* (*Part-Of-Speech tagging*) est effectué via *TreeTagger* [Schmid 1997] : cela consiste à associer aux mots d'un texte les informations grammaticales correspondantes (en rouge dans la figure 7).
3. suivi d'un étiquetage des termes via *TermTagger* utilisé avec les ressources mentionnées précédemment : cela consiste à identifier les termes présents dans nos ressources, ils sont étiquetés grammaticalement "catSem" et on ajoute la catégorie sémantique (en violet dans la figure 7).

Un exemple de format en sortie de ce pré-traitement est présenté à la figure 7 :

Through/IN/through// the/DT/the// inhibition/NN/inhibition// of/IN/of// this/DT/this// enzyme/catSem/Enzyme /drug/ system/NN/system// ,/,/,// grapefruit/catSem/grapefruit/food/ juice/catSem/Juice/drug/ interacts/VBZ/interact// with/IN/with// a/DT/a// variety/NN/variety// of/IN/of// medications/NNS/medication// ,/,/,// leading/VBG/lead// to/TO/to// elevation/NN/elevation// of/IN/of// their/PP/their// serum/NN/serum// concentrations/NNS/concentration// ./SENT/.// The/DT/the// effects/NNS/effect// of/IN/of// ginkgo/NN/ginkgo/food/ and/CC/and// garlic/catSem/garlic/food/ on/IN/on// warfarin/catSem/Warfarin /drug/ use/NN/use// ./SENT/.// [...] effects/NNS/effect// of/IN/of// grapefruit/catSem/grapefruit/food/ juice/catSem/Juice/food/ on/IN/on// sertraline/catSem/Sertraline /drug/ [...]

FIGURE 7 – Exemple du corpus taggé après pré-traitement

À cela nous pouvons ajouter des informations d'indexation afin d'identifier un mot dans tout le corpus et ainsi pouvoir y référer lorsqu'une relation y est liée.

6 Expérimentations et Résultats

6.1 Expérimentations

Dans le cadre de la première initialisation en contexte, nous proposons d'extraire nos premières relations à l'aide des patrons suivants :

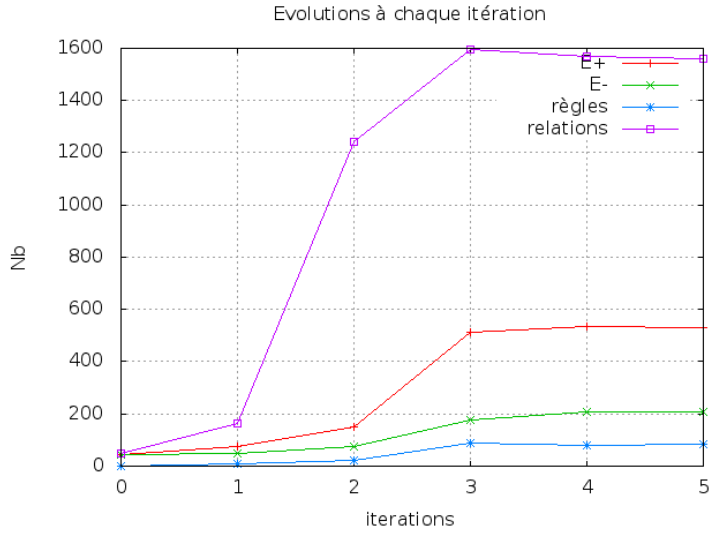
- effect of FOOD on DRUG (60 occurrences)
- impact of FOOD on DRUG (10 occurrences)
- FOOD increase DRUG (50 occurrences)
- FOOD inhibits DRUG (20 occurrences)

Ces derniers sont appliqués sur notre corpus duquel 140 exemples (vérifiés à la main) ont pu être extraits. Ces exemples sont générés avec une classe *relation*, *increase* ou *decrease*.

On considèrera que tout couple aliment/médicament dans une phrase, n'ayant pas été sélectionné par les patrons précédents, est un exemple négatif (méthode décrite en section 4.3). On verra que ce choix un peu strict va empêcher ce cycle d'apprentissage de découvrir de nouveaux contextes (pour le moment, nous n'avons malheureusement pas trouvé de solution pour améliorer cette étape). Dans un premier temps, le but de ces tests était de retrouver les patrons utilisés lors de l'initialisation et de vérifier la faisabilité de l'approche.

6.2 Résultats

N'ayant pas encore pu annoter le corpus afin d'établir des données de test, nous allons essentiellement analyser le nombre de relations et règles générées à chaque itération :



Itération	E+/E-	Nb Règles	Nb Relations
0	42/46	2	48
1	74/47	9	165
2	152/77	20	1244
3	511/175	90	1594
4	535/209	81	1568
5	532/206	82	1559

On peut observer une certaine stabilisation dès la troisième itération. Nous constatons une augmentation de 70 règles entre l'itération 2 et 3 mais qui ne se ressent pas sur le nombre de relations (augmentation de 300 relations seulement). Nous pouvons penser que ces nouveaux patrons appris décrivent des contextes similaires ou déjà connus et donc n'apportent pas plus de relations. Cependant ces patrons semblent s'affiner comme nous pouvons le voir entre l'itération 4 et 5 (perte de 9 règles).

Il a fallu trouver la configuration optimale d'Aleph pour obtenir des résultats convenables (voir section 4.5). Les règles trouvées étaient souvent trop génériques, elles manquaient de précisions. De plus, les exemples sont générés selon des patrons que nous attendions en retour. Les résultats les plus proches sont les suivants :

- Règle 1 : Couverture = 67
  rel(A,B,relation) :-
  pred(A,C) ∧ pred(C,D) ∧ has_lemma(D,effect) ∧ has_lemma(C,of).
  D ← C ← A
  effect of [catSem]

- Règle 2 : Couverture = 19
  rel(A,B,relation) :-
  pred(B,C) ∧ has_lemma(C,on).
  C ← B
  on [catSem]

- Règle 3 : Couverture = 41
  rel(A,B,relation) :-
  has_lemma(C,of) ∧ pred(B,C) ∧ pred(A,D).
  C ← B D ← A
  of [catSem]... [catSem]

- Et si nous combinons les régles 1 et 2 on peut obtenir :
  effect of [catSem] on [catSem]

Nous pouvons constater que la précision et la couverture de certaines règles sont encore faibles, on peut supposer que cela est dû au manque de flexibilité sur la distance entre les mots. En effet, *Aleph* sélectionne une règle en fonction de sa couverture sur les exemples sélectionnés. Ici on distingue clairement le patron "effect of FOOD on DRUG" cependant le corpus présente surtout des contextes

de la forme : "effect of [...]* [food] [...]* on [...]* [drug] [...]*". Il y a donc souvent plusieurs mots autour des termes. C'est pour cette raison que nous observons des patrons partiels [Saeger et al. 2009].

L'idéal serait de disposer de la notion de distance afin de trouver une règle telle que :

rel(A,B,relation) :-
pred(C,D), has_lemma(D,effect), has_lemma(C,of), has_lemma(E,on),
distance(C,A,CA), CA $\leq$ 3,
distance(A,E,AE), AE $\leq$ 3,
distance(E,B,EB), EB $\leq$ 4.

Cependant le calcul devient excessivement long lorsqu'on détermine une distance entre chaque mot de phrase (cf section 4.2).

Tout cela n'a pas empêché la méthode de découvrir des contextes nouveaux (pour certains, proches de ceux donnés initialement) :

[food] \w+ inhibit(s) \w+ [drug]
The grapefruit juice inhibits a crucial enzyme

of [food] (IN) \w+ pharmacokinetics \w+ [drug]
impact of grapefruit juice and seville orange juice on the pharmacokinetics of dextromethorphan

the \w+ of [food] (IN) [drug]
The effect of pineapple juice on the expression of CYP1A1, CYP2E1

(IN) [food] \w+ [drugeffect] \w+ [drug]
Effect of food on the oral bioavailability of didanosine

FIGURE 8 – Patrons découverts

On peut constater l'introduction de nouvelles ancres lexicales telles que *inhibits* et *pharmacokinetics*. On peut également observer un type sémantique non introduit jusqu'à présent (*drugeffect*), ceci étant pour montrer qu'il est rapidement possible d'obtenir des résultats plus précis en enrichissant le corpus de nouvelles connaissances sémantiques.

Une sortie Brat[Stenetorp et al. 2012] [8] est également générée dans le but de mieux observer les résultats, pouvoir vérifier, corriger et annoter plus facilement le corpus d'apprentissage (voir la capture d'écran 6.2).

8. http ://brat.nlplab.org/

7 Conclusions et Perspectives

7.1 Conclusions

Nous avons défini une méthode d'apprentissage de patrons lexico-sémantiques utilisant la PLI. Les premiers résultats sont encourageant mais nécessitent des améliorations. Nous avons constaté que le paramétrage doit être précisément défini en fonction des besoins et du corpus.

Nous avons vu qu'il est possible de choisir différents types d'amorçage en fonction de l'objectif final (cf 4.3) :

- Si le but est d'affiner certains patrons, l'initialisation à partir de patrons connus avec vérifications des relations sera la meilleure approche.
- Si le but est de découvrir de nouveaux contextes propices, alors il faudra préférer l'acquisition (manuelle ou à l'aide de ressources) d'exemples en contexte à partir des relations connues. La technique du *One vs All* sera également conseillée. (cette méthode est en cours d'implémentation)
- Si le but est d'acquérir de nouvelles relations, il est bien évidemment déconseillé de vérifier si les relations extraites sont présentes dans les ressources terminologiques.

7.2 Perspectives

Afin de mieux évaluer la méthode il est nécessaire de comparer les résultats obtenus après utilisation des patrons générés sur le corpus de test. Pour ce faire, il faut annoter le corpus, ce travail est en cours.
Une perspective consiste à aller plus loin encore dans les paramétrages et adaptation des fonctions de scores.

De nouvelles pistes de recherche sont apparues concernant le post-traitement et l'adatation des paramètres en fonction des données :

- L'observation des intersections de couverture des règles et les possibilités de fusion des patrons partiels
- L'adaptabilité des paramètres (bornes mentionnées dans la section 4.5) en fonction du nombre de relations et de l'hétérogénéité du corpus.

Certaines tâches sont encore à faire comme l'amélioration de pré-traitement :

- Nettoyages de ressources terminologiques.
- Implémentation de l'initialisation 2 mentionnée en section 4.

Il est également intéressant d'étudier plus en détail les méthodes hybrides comme les *Markov Logic Network* (MLN) ainsi que les outils utilisés : Alchemy et thebeast [Riedel 2008]. En effet, il pourrait être intéressant d'introduire des probabilités sur chacune des clauses et contraintes pour obtenir des règles pondérées.


relation
Food
modeadm
Drug Effect
Drug
2 Clinical studies in humans showed that fruit juices reduced the oral bioavailability of fexofenadine by preferentially inhibiting OATP over P-gp .
relation
Food
modeadm
Drug
Food
Side Effect
3 The objective of this study was to investigate the effects of fruit juices on the oral absorption of fexofenadine in rats to establish a preclinical fruit juice- drug interaction model .
Drug
Other
4 In rats , fexofenadine was excreted unchanged in the urine , bile , and gastrointestinal tract , indicating minimal metabolism , making it an ideal probe to study transport processes .
Drug
Drug
modeadm
Drug
dosage
5 Coadministration of fexofenadine with ketoconazole , a P-gp inhibitor , increased the oral exposure of fexofenadine by 187% 187%. .
relation
Drug
Food
Food
modeadm
Drug
numbers
dosage
6 In contrast , coadministration of fexofenadine with orange juice or apple juice to rats decreased the oral exposure of fexofenadine by 31 and 22% , respectively .
relation
Food
Food
modeadm
Drug
numbers
dosage
7 Increasing the quantity of orange or apple juice administered further decreased the oral exposure of fexofenadine , by 40 and 28% , respectively .
Drug
Drug Effect
8 This reduction in fexofenadine bioavailability was moderate compared to that seen in humans .
relation
Food
Drug
modeadm
9 These findings suggest that in rats fruit juices may also preferentially inhibit OATP rather than P-gp-mediated transport in fexofenadine oral absorption , albeit to a lesser extent
Food
Side Effect
Food
Side Effect
10 This fruit juice juice- -drug interaction rat model may be useful in prediction of potential food food- -drug interactions in humans for drug candidates .
numbers
11 Copyright 2004 Wiley-Liss , Inc. .
relation
Food
food2
drug2
12 The grapefruit challenge: the juice inhibits a crucial enzyme , with possibly fatal consequences .

Références

[ACL 1993] Manning (Christopher D.). – Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. *In : Proceedings of the 31st Annual Meeting of the ACL.* – Morristown, NJ : Association for Computational Linguistics.

[Bailey et al. 1991] Bailey (Dg), Spence (Jd), Munoz (C.) et Arnold (J.M.O.). – Interaction of citrus juices with felodipine and nifedipine. *In : The Lancet Volume 337.* pp. 268–269. – The Lancet.

[Bodenreider et al. 2001] Bodenreider (Olivier), Burgun (Anita) et Rindflesch (Thomas). – Lexically-suggested Hyponymic Relations among Medical Terms and their Representation in the UMLS. *In : TIA 2001*, pp. 11–21. – Nancy, France, 2001.

[Church et Hanks 1990] Church (Kenneth W.) et Hanks (Patrick). – Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol. 16 (1), March 1990, pp. 22–29.

[Claveau et Sébillot 2004] Claveau (Vincent) et Sébillot (Pascale). – Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. *In : Actes de la conférence TALN'2004.* – Fès, Maroc, avril 2004.

[Claveau 2003] Claveau (Vincent). – *Acquisition automatique de lexiques sémantiques pour la recherche d'information.* – Thèse de doctorat, Université de Rennes 1, décembre 2003.

[Dunning 1993] Dunning (Ted). – Accurate Methods for the Statistics of Suprise and Coincidence. *Computational Linguistics*, vol. 19 (1), march 1993, pp. 61–74. – Special Issue on Using Large Corpora : I.

[Grabar et al. 2004] Grabar (Natalia), Malaisé (Véronique), Marcus (Aurélia) et Krul (Alexandra). – Repérage de relations terminologiques transversales en corpus. *In : Actes de la conférence TALN'2004.* – Fès, Maroc, avril 2004.

[Grabar et Zweigenbaum 1999] Grabar (Natalia) et Zweigenbaum (Pierre). – Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. *In : Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 1999)*, pp. 175–184. – Cargèse, France, 1999.

[Grabar et Zweigenbaum 2003] Grabar (Natalia) et Zweigenbaum (Pierre). – Lexically-Based Terminology Structuring. *Unknown*, vol. 10, 2003, pp. 23–54.

[Grefenstette 1994] Grefenstette (Gregory). – *Exploration in Automatic Thesaurus Discovery.* – Boston, USA, Kluwer Academic Publishers, 1994.

[Habert et Nazarenko 1996] Habert (Benoît) et Nazarenko (Adeline). – La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience. *In : Actes des Journées Acquisitions des Connaissances*, pp. 137 – 142. – Sète, France, May 1996.

[Hamon et al. 2010a] Hamon (Thierry), Graña (Martin), Raggio (Víctor), Grabar (Natalia) et Naya (Hugo). – Identification of relations between risk factors and their pathologies or health conditions by mining scientific literature. *In : Proceedings of MEDINFO 2010*, pp. 964–968. – PMID : 20841827.

[Hamon et al. 2010b] Hamon (Thierry), Périnet (Amandine), Nobécourt (Jérôme) et Grabar (Natalia). – Linguistic and semantic annotation for information extraction and characterization. *In : Proceedings of the I2B2 WorkshopS*.

[Hamon 1998] Hamon (Thierry). – Does the general semantic information help the terminology structuration ? *In : Proceedings of the Workshop on Lexical Semantic Systems (WLSS'98)*. – Pise, Italie, 1998.

[Harris 1971] Harris (Zellig S.). – *Structures mathématiques du langage*. – Paris, Monographies de linguistique mathématique. Dunod, 1971.

[Hearst 1992] Hearst (Marti A.). – Automatic Acquisition of Hyponyms from Large Text Corpora. *In : Proceedings of 14th International Conference on Computational Linguistics (COLING'92)*, pp. 539–545. – Nantes, France, August 1992.

[IS 2005] Ibekwe-SanJuan (Fidelia). – Inclusion lexicale et proximité sémantique entre termes. *In : Actes de la conférence TIA 2005*, pp. 45–57. – Rouen, avril 2005.

[Jacquemin et al. 1997] Jacquemin (Christian), Klavans (Judith L.) et Tzoukermann (Evelyne). – Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. *In : Proceedings of the ACL'97/EACL'97*, pp. 24–31. – Barcelona, Spain, 1997.

[Lafferty et al. 2001] Lafferty (John D.), McCallum (Andrew) et Pereira (Fernando C. N.). – Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. *In : Proceedings of the Eighteenth International Conference on Machine Learning*. pp. 282–289. – San Francisco, CA, USA, 2001.

[Martienne et Morin 1999] Martienne (Emmanuelle) et Morin (Emmanuel). – *Using s Symbolic Machine Learning Tool to Refine Lexico-syntactic Patterns*. – Rapport de Recherche n° 183, Institut de Recherche en Informatique de Nantes (IRIN), Avril 1999.

[Mccrae et Collier 2008] Mccrae (John) et Collier (Nigel). – Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics*, vol. 9, March 2008, pp. 159+.

[Meyer 2001] Meyer (I.). – Extracting Knowledge-rich Contexts for Terminography. *In : Recent Advances in Computational Terminology*, éd. par Bourigault (D.), Jacquemin (C.) et L'Homme (M.C.). pp. 279–302. – Amsterdam/Philadelphia, 2001.

[MH 2013] Morlane-Hondère (François). – *Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique*. – These, Université Toulouse le Mirail - Toulouse II, juillet 2013.

[Minard 2012] Minard (Anne-Lyse). – *Extraction d'information multi-documents en domaine de spécialité.* – Thèse pour l'obtention du diplôme d'état de docteur en informatique, Université Paris Sud, December 2012.

[Morin et Martienne 1999] Morin (Emmanuel) et Martienne (Emmanuelle). – Raffinement de patrons lexico-syntaxiques. *In : Actes de IC'99 (Ingenierie des Connaissances)*. Plate-forme AFIA, pp. 89–96. – Palaiseau, France, juin 1999.

[Morin 1998] Morin (Emmanuel). – Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes. *In : Actes de la Conférence TALN 1998*, pp. 172–181. – Paris, France, 1998.

[Morin 1999] Morin (Emmanuel). – Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues*, vol. 40 (1), 1999.

[Muggleton et Raedt 1994] Muggleton (Stephen) et Raedt (Luc De). – Inductive Logic Programming : Theory and Methods. *JOURNAL OF LOGIC PROGRAMMING*, vol. 19 (20), 1994, pp. 629–679.

[National Library of Medicine 2003] National Library of Medicine (édité par). – *UMLS Knowledge Source.* – NLM, 2003, 13^{th} édition.

[Richardson et Domingos 2006] Richardson (Matthew) et Domingos (Pedro). – Markov Logic Networks. *Mach. Learn.*, vol. 62 (1-2), février 2006, pp. 107–136.

[Riedel 2008] Riedel (Sebastian). – Improving the accuracy and Efficiency of MAP Inference for Markov Logic. *In : Proceedings of the 24th Annual Conference on Uncertainty in AI (UAI '08)*, pp. 468–475.

[Roberts et al. 2008] Roberts (Angus), Gaizauskas (Robert), Hepple (Mark) et Guo (Yikun). – Mining clinical relationships from patient narratives. *BMC Bioinformatics*, vol. 9 (Suppl 11), 2008, p. S3.

[Rosario et Hearst 2004] Rosario (Barbara) et Hearst (Marti A.). – Classifying semantic relations in bioscience texts. *In : In Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (Datasets : http ://biotext.berkeley.edu/data.html)*, pp. 431–438.

[RxNorm 2009] Unknown . – *RxNorm, a standardized nomenclature for clinical drugs.* – Rapport technique, Bethesda, Maryland, National Library of Medicine, 2009. Available at `www.nlm.nih.gov/research/umls/rxnorm/docs/index.html`.

[Røst et al. 2010] Røst (Thomas B.), Akbar (Saiful), Øystein Nytrø et Basgalupp (Márcio). – Medical Relation Extraction with Semantic Grammars. *In : Proceedings of the workshop I2B2 2010.*

[Saeger et al. 2009] Saeger (Stijn De), Torisawa (Kentaro), Kazama (Jun'ichi), Kuroda (Kow) et Murata (Masaki). – Large Scale Relation Acquisition Using Class Dependent Patterns. *2013 IEEE 13th International Conference on Data Mining*, vol. 0, 2009, pp. 764–769.

[Sahay et al. 2008] Sahay (Saurav), J. (Lee) et N. (Krishnamurthi). – Relationship Extraction from Biomedical Documents using Conditional Random Fields. *In : Relationship Extraction from Biomedical Documents using Conditional Random Fields*, p. 43.

[Schmid 1997] Schmid (Helmut). – Probabilistic Part-of-Speech Tagging Using Decision Trees. *In : New Methods in Language Processing Studies in Computational Linguistics*, éd. par Jones (Daniel) et Somers (Harold).

[Schumann 2011] Schumann (Anne-Kathrin). – A Case Study of Knowledge-Rich Context Extraction in Russian. *In : Proceedings of the 9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, éd. par Kageura (Kyo) et Zweigenbaum (Pierre), pp. 143–146. – INaLCO, 2011.

[Smadja 1993] Smadja (Franck). – Retrieving Collocations from Text : Xtract. *Computational Linguistics*, vol. 19 (1), march 1993, pp. 143–177. – Special Issue on Using Large Corpora : I.

[Stenetorp et al. 2012] Stenetorp (Pontus), Pyysalo (Sampo), Topić (Goran), Ohta (Tomoko), Ananiadou (Sophia) et Tsujii (Jun'ichi). – 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*. pp. 102–107. – Avignon, France, April 2012.

[Uzuner et al. 2010] Uzuner (Özlem), Solti (Imre) et Cadag (Eithon). – Extracting medication information from clinical text. *JAMIA*, vol. 17 (5), 2010, pp. 514–518.

[VN 2013] Vivi Nastase , Preslav Nakov (Diarmuid Ó Séaghdha et Stan Szpakowicz). – *Semantic Relations Between Nominals*. – Morgan and Claypool Publishers, 2013.

[Weissenbacher 2004] Weissenbacher (Davy). – La relation de synonymie en génomique. *In : Actes de la conférence RECITAL'2004*, pp. 298–303. – Fès, Maroc, avril 2004.

[Wishart et al. 2006] Wishart (David S.), Knox (Craig), Guo (An Chi), Shrivastava (Savita), Hassanali (Murtaza), Stothard (Paul), Chang (Zhan) et Woolsey (Jennifer). – DrugBank : a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, vol. 34, 2006, p. D668–D672. – Database issue.