Une expérience d'extraction de relations sémantiques à partir de textes dans le domaine médical

Mehdi EMBAREK, Olivier FERRET
CEA LIST, LIC2M,
18 route du Panorama, BP6, FONTENAY AUX ROSES, F- 92265 France
{embarekm,ferreto}@zoe.cea.fr

Résumé. Dans cet article, nous présentons une méthode permettant d'extraire à partir de textes des relations sémantiques dans le domaine médical en utilisant des patrons linguistiques. La première partie de cette méthode consiste à identifier les entités entre lesquelles les relations visées interviennent, en l'occurrence les maladies, les examens, les médicaments et les symptômes. La présence d'une des relations sémantiques visées dans les phrases contenant un couple de ces entités est ensuite validée par l'application de patrons linguistiques préalablement appris de manière automatique à partir d'un corpus annoté. Nous rendons compte de l'évaluation de cette méthode sur un corpus en Français pour quatre relations.

Abstract. In this article, we present a method to extract semantic relations automatically in the medical domain using linguistic patterns. This method consists first in identifying the entities that are part of the relations to extract, that is to say diseases, exams, treatments, drugs and symptoms. Thereafter, sentences that contain these entities are extracted and the presence of a semantic relation is validated by applying linguistic patterns that were automatically learnt from an annotated corpus. We report the results of an evaluation of our extraction method on a French corpus for four relations.

Mots-clés : extraction de relations sémantiques, patrons lexico-syntaxiques, domaine médical.

Keywords: extraction of semantic relations, lexico-syntactic patterns, medical domain.

1 Introduction

Dans cet article, nous nous intéressons au domaine médical, dont la particularité est la richesse et la complexité de son vocabulaire spécialisé. Cette particularité a conduit depuis de nombreuses années au développement d'un ensemble important de ressources terminologiques telles que le MeSH ou l'UMLS par exemple. Ces ressources ont été utilisées dans des contextes aussi divers que l'indexation de documents, la recherche d'information, l'extraction d'information ou même les systèmes de question-réponse. À l'image de réseaux lexicaux de même type mais plus généraux, comme WordNet (Fellbaum, 1998), ces ressources contiennent majoritairement des relations d'hyperonymie ou de synonymie et sont donc beaucoup moins riches en relations syntagmatiques comme celles spécifiant qu'une maladie *M* peut être soignée par le traitement *T* ou que l'examen *E* permet de diagnostiquer la maladie *M*. De même, les méthodes ayant pour

objectif d'extraire des relations sémantiques à partir de textes se focalisent majoritairement sur les relations de synonymie et d'hyperonymie, à la suite de (Hearst, 1992) ou plus récemment de (Caraballo, 1999).

Le travail dont nous rendons compte dans cet article concerne l'extraction et la validation de relations sémantiques entre des entités caractéristiques du domaine médical, telles que des maladies, des médicaments ou des examens, en se focalisant prioritairement sur des relations de type syntagmatique. Différents travaux ont déjà été menés concernant l'extraction de relations sémantiques dans le domaine médical ou biomédical, travaux parmi lesquels on peut citer (Craven, 1999), (Mukherjea & Sahay, 2006), (Rosario & Hearst, 2004) ou (Vintar & Buitelaar, 2003). Les recherches menées en extraction d'information dans ce même contexte, bien qu'ayant *a priori* une finalité plus large, se ramènent dans bon nombre de cas à l'extraction de ce même type de relations, à l'instar de la détection des interactions entre gènes ou entre gènes et protéines. On se reportera à (Nédellec, 2004) pour un panorama de ces travaux, souvent fondés sur des règles d'extraction définies manuellement.

La méthode que nous proposons repose pour sa part sur l'identification puis l'application de patrons linguistiques caractérisant les relations visées, dans le prolongement direct de (Pantel *et al*., 2004). Cette application se déroule en deux étapes. La première consiste à identifier dans les textes les entités du domaine médical intervenant dans les relations visées. Dans la phrase « ... en novembre 2001, année d'un cancer de la prostate traité par radiothérapie et qu'il affirme aujourd'hui disparu, ... », le premier objectif est ainsi de repérer que *cancer de la prostate* est une maladie et que *radiothérapie* est un traitement. Dans un second temps, l'application du patron <*maladie*> *traité par* <*traitement*> construit automatiquement à partir d'un corpus de référence permet de valider la présence d'une relation entre ces deux entités, relation stipulant dans le cas présent que la *radiothérapie* est un traitement possible du *cancer de la prostate*.

2 Ontologie du domaine médical

La première étape de notre travail s'est focalisée sur la définition d'une ontologie du domaine de la médecine générale permettant de faire apparaître les entités caractérisant ce domaine ainsi que les relations existant entre ces entités. Cette ontologie a été définie à la fois en sollicitant directement des médecins et par l'analyse des questions typiquement posées par des médecins généralistes (Ely *et al*., 1999). La Figure 1 illustre le sous-ensemble de cette ontologie correspondant aux quatre relations objets de notre travail et aux entités qu'elles font intervenir. Sont

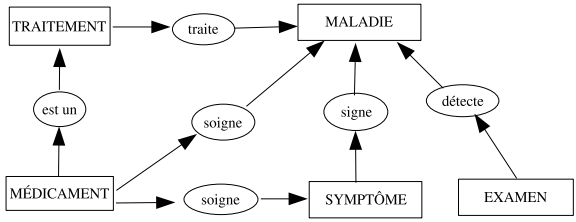

FIG. 1 – Sous-ensemble de l'ontologie du domaine médical concernant les relations à extraire

ainsi concernées les cinq entités suivantes : Maladie, Traitement, Médicament, Symptôme et Examen. Parmi toutes les relations de la Figure 1, les quatre relations auxquelles nous nous sommes attachés sont :

– Traite : Maladie – Traitement
– Soigne : Maladie – Médicament
– Détecte : Maladie – Examen
– Signe : Maladie – Symptôme

3 Reconnaissance des entités médicales

L'identification dans les documents des entités médicales que nous avons retenues a été réalisée indépendamment de leur nature intrinsèque (entité nommée au sens littéral ou terme) en adoptant une approche à base de règles mêlant patrons morpho-syntaxiques et listes d'entités ou d'éléments caractéristiques de ces entités. Ces règles ont été définies manuellement à partir d'un travail sur corpus. Nous avons repris en l'occurrence une des approches classiquement utilisées pour identifier des entités nommées de nature plus générale comme les personnes, les organisations ou les lieux. À la différence de ces dernières, les patrons morpho-syntaxiques ont ici une importance moindre ce qui à l'inverse, donne un rôle plus central aux listes d'entités ou de parties d'entités. Nous avons donc accordé un soin tout particulier à la constitution de ces listes pour chaque type d'entités en exploitant pour ce faire à la fois des ressources disponibles sur le Web[1] et des dictionnaires de l'Académie de Médecine sous forme électronique.

Chaque règle de reconnaissance d'une entité est composée d'un déclencheur, d'un contexte précédent, d'un contexte suivant et du type d'entité identifié. Ces règles sont implémentées sous la forme d'automates. Leur application s'effectuant à la suite de l'étiquetage morpho-syntaxique réalisé par l'analyseur LIMA (LIC2M Multilingual Analyzer) (Besançon & Chalendar, 2005), le déclencheur et les contextes précédent et suivant d'une règle prennent donc la forme d'expressions régulières pouvant porter sur la forme fléchie, la forme normalisée ou la catégorie d'un ou de plusieurs mots. Ainsi, la règle

@AnnonceurMaladie::$L_DET ?::$L_DET ($L_NC|$L_NP)::MALADIE[2]
déclencheur::*contexte_précédent*::*contexte_suivant*::*type_d'expression*

permet-elle d'identifier *maladie de Lyme* comme une maladie dans « La maladie de Lyme est une … » tandis que la règle

[@AnnonceurSymptome]::::[,] [$L_NC] [$L_DET] $L_NC::SYMPTOME[3]

reconnaît *fièvre* comme un symptôme dans « … symptôme, comme la fièvre … ». On peut noter à cette occasion la présence de référence à des listes permettant de regrouper des éléments linguistiques ayant un même rôle, comme les éléments marquant la présence d'une maladie (*@AnnonceurMaladie={maladie, syndrome …}*) ou ceux marquant la présence d'un symptôme (*@AnnonceurSymptome={signe, symptôme …}*).

[1] Le site Doctissimo pour les noms de médicaments ou le site Orphanet pour les noms de maladies par exemple.

[2] ? marque classiquement un élément optionnel tandis que (|) note une alternative. *$L_DET*, *$L_NC* et *$L_NP* sont des catégories morpho-syntaxiques, correspondant respectivement à déterminant, nom commun et nom propre.

[3] [] permet de spécifier la non appartenance d'un élément à l'entité reconnue.

4 Extraction de relations sémantiques

4.1 Apprentissage de patrons linguistiques d'extraction

Le terme de patron linguistique désigne dans le cas présent un schéma lexico-syntaxique spécifique d'une relation sémantique intervenant entre deux entités. Dans le cas présent, ces patrons sont dits multi-niveaux, c'est-à-dire qu'ils s'appuient sur des informations provenant de plusieurs niveaux de traitement des textes : à l'instar des règles de reconnaissance des entités médicales, ils peuvent ainsi faire intervenir la forme fléchie des mots, leur forme normalisée ou bien encore leur catégorie morpho-syntaxique. Le processus que nous avons élaboré pour extraire à partir d'un corpus les patrons linguistiques caractérisant une relation est le suivant :

1. appliquer sur le corpus considéré les règles de reconnaissance des entités médicales impliquées dans la relation cible. Nous prendrons à titre d'exemple la relation *Traite* entre une *Maladie* et un *Traitement* ;
2. extraire du corpus toutes les phrases contenant les deux entités de la relation cible, à savoir ici les phrases contenant à la fois une maladie et un traitement ;
3. sélectionner manuellement les phrases dans lesquelles la relation entre les deux entités correspond effectivement à la relation cible. Cela implique en particulier d'écarter les phrases telles que « la <maladie> n'est pas traitée par le <traitement> » ;
4. réaliser l'analyse linguistique de chaque phrase sélectionnée pour en faire apparaître les différents niveaux d'information. Cette analyse est réalisée comme pour la reconnaissance des entités par l'analyseur LIMA ;
5. remplacer dans chaque phrase les entités par leur type ;
6. appliquer l'algorithme d'extraction de patrons multi-niveaux présenté ci-dessous entre chaque couple de phrases parmi celles sélectionnées précédemment.

Pour extraire les patrons linguistiques propres à chaque relation sémantique traitée (cf. Section 2), nous faisons appel à l'algorithme proposé par Ravichandran dans (Pantel *et al.*, 2004) pour apprendre des patrons multi-niveaux. Cet algorithme est composé de deux parties. La première consiste à calculer la distance d'édition minimale entre deux phrases, ce qui permet de déterminer le nombre minimum d'opérations (insertion, suppression et remplacement) à appliquer pour passer d'une phrase à l'autre. La deuxième étape extrait le patron multi-niveau le plus spécifique permettant de généraliser les deux phrases. Pour compléter certains alignements, deux opérateurs sont introduits : *(*s*)*, qui représente 0 ou 1 instance de n'importe quel mot et *(*g*)*, qui représente exactement une instance de n'importe quel mot.

Nous avons appliqué le processus décrit plus haut sur une partie du corpus médical de la campagne d'évaluation des systèmes de Question/Réponse EQueR et extrait ainsi des patrons multi-niveaux pour les quatre relations considérées dans cette étude. Nous donnons à titre illustratif quelques exemples de patrons extraits pour chaque relation :

Maladie – Examen

<examen> en suspicion de <maladie>
<maladie> être *(*g*)* à le <examen>
<examen> pour le NC_GEN *(*g*)* <maladie>
<examen> montre un <maladie>

<examen> *(*g*)* le diagnostic *(*g*)* <maladie>
<maladie> , *(*s*)* <examen>

Maladie – Traitement
<traitement> dans le traitement des <maladie>
<traitement> être *(*g*)* PREP_GENERAL le traitement de le *(*s*)* <maladie>
<traitement> est recommandé pour le traitement des <maladie>
<maladie> , *(*g*)* NC_GEN *(*g*)* une <traitement>
<maladie> nécessitant un <traitement>
<traitement> contre le <maladie>

Maladie – Symptôme
<maladie> , se manifeste par une <symptome>
<symptome> *(*g*)* être *(*s*)* des symptômes d' une <maladie>
<maladie> VERBE_PRINC_INDICATIF *(*s*)* <symptome>
<symptome> *(*s*)* peut VERBE_PRINC_INFINIT la NC_GEN de le *(*s*)* <maladie>
<maladie> *(*g*)* avec <symptome>
<symptom> (<maladie>

Maladie – Médicament
<medicament> est indiqué dans le traitement de la <maladie>
<medicament> , utilisée *(*s*) (*s*)* dans le traitement de <maladie>
<medicament> est un médicament utilisé pour traiter <maladie>
<maladie> chez les NC_GEN traité par <medicament>
<medicament> dans le cas de <maladie>
<medicament> (proposé dans le traitement de *(*s*)* <maladie>

4.2 Extraction et validation des relations sémantiques

Pour acquérir de nouvelles relations sémantiques à partir d'un corpus, *i.e.* de nouveaux couples d'entités liées par une relation identifiée, nous appliquons une démarche en deux temps. Comme dans le cas de l'extraction des patrons de relation, nous commençons par sélectionner des relations candidates en repérant les phrases contenant un couple d'entités intervenant dans une des relations cibles. Dans un second temps, nous confrontons la phrase contenant la relation candidate avec les patrons linguistiques spécifiques de cette relation. Si l'un au moins de ces patrons peut s'appliquer à la phrase considérée, la relation est considérée comme validée. Dans le cas contraire, elle est écartée. Plus formellement, le processus mis en œuvre pour un type de relation est le suivant :

1. appliquer sur le corpus considéré les règles de reconnaissance des entités médicales impliquées dans la relation cible ;
2. extraire du corpus toutes les phrases contenant simultanément les deux entités de la relation cible ;
3. réaliser l'analyse linguistique de chaque phrase sélectionnée, toujours en utilisant l'analyseur LIMA ;

4. remplacer dans chaque phrase les entités par leur type ;
5. pour chaque phrase, calculer sa distance d'édition avec tous les patrons multi-niveaux de la relation. Si la distance d'édition est égale à 0, c'est-à-dire si la relation entre les deux types sémantiques de la phrase respecte le schéma du patron, alors valider la relation.

Nous avons appliqué notre algorithme d'extraction et de validation de relations sémantiques pour les quatre relations retenues dans notre étude sur un corpus de textes médicaux recueillis dans le cadre du projet Technolangue Atonant, corpus différent de celui utilisé pour induire les patrons de caractérisation de ces relations. Voici quelques exemples de relations sémantiques validées par notre méthode (le patron utilisé est introduit par $\Longrightarrow$) :

Maladie – Examen
tomodensitométrie dans le diagnostic des *tumeurs du médiastin*
$\Longrightarrow$ <examen> dans le diagnostic *(*s*)* <maladie>
radiographie pulmonaire pour le diagnostic de *tuberculose*
$\Longrightarrow$ <examen> *(*g*)* le diagnostic *(*g*)* <maladie>

Maladie – Médicament
insuffisance rénale chronique traitée par *Eprex*
$\Longrightarrow$ <maladie> traitée par <medicament>
Le *vaccin* utilisé pour prévenir la *fièvre aphteuse*
$\Longrightarrow$ <medicament> utilisé pour VERBE_PRINC_INFINIT *(*g*)* <maladie>

Maladie – Traitement
chimioprophylaxie contre la *malaria*
$\Longrightarrow$ <traitement> contre la <maladie>
radiothérapie dans le traitement de la *resténose*
$\Longrightarrow$ <traitement> dans le traitement de la <maladie>

Maladie – Symptôme
L'*intoxication* peut provoquer des *vomissements*
$\Longrightarrow$ <maladie> *(*s*)* peut VERBE_PRINC_INFINIT DET_ART_CONTRACT *(*s*)* <symptome>
Botulisme , se manifeste par une *sécheresse de la bouche*.
$\Longrightarrow$ <maladie>, se manifeste *(*s*)* par une <symptome>

5 Évaluation

Dans cette section, nous présentons les résultats des évaluations menées sur deux corpus en français, constitués chacun d'articles scientifiques et de recommandations de bonne pratique médicale téléchargées à partir du site du CISMeF[4]. La première (cf. Table 1) concerne l'identification des entités médicales dans les textes en appliquant les règles de reconnaissance présentées à la Section 3. La seconde (cf. Table 2) porte sur l'extraction et la validation de relations sémantiques grâce à la méthode présentée à la Section 4.2.

La Table 1 résume les résultats obtenus en appliquant nos règles de reconnaissance d'entités

[4]Catalogue et Index des Sites Médicaux Francophones : `http://www.cismef.org`

médicales sur un sous-ensemble d'une taille de 1,5 Mo (soit environ 130.000 mots) du corpus médical de la campagne d'évaluation EQueR. Les mesures utilisées sont classiquement la précision et le rappel, qui se définissent ici de la façon suivante :
– la précision représente le nombre d'entités correctes extraites par notre système sur le nombre total des entités extraites par notre système ;
– le rappel représente le nombre d'entités correctes extraites par notre système sur le nombre total des entités présentes dans le corpus.
La F1-mesure correspond à la moyenne harmonique entre la précision et le rappel. Ces mesures sont réalisées par comparaison avec une annotation manuelle du corpus d'évaluation. Les résul-

Entités sémantiques	Précision	Rappel	F1-mesure
Maladie	0,95	0,80	0,86
Symptôme	0,84	0,76	0,79
Examen	0,94	0,93	0,93
Traitement	0,86	0,81	0,83
Médicament	0,93	0,88	0,90
Moyenne	0,90	0,84	0,86

TAB. 1 – Résultats de la reconnaissance des entités médicales

tats de notre méthode donnés par la Table 1 montrent une précision et un rappel supérieurs ou égaux à 83% en moyenne, ce qui constitue un bon niveau pour ce type de tâche. On peut noter en particulier le niveau élevé de la précision qui caractérise un niveau de fiabilité très significatif. Cette propriété est d'autant plus importante dans le cas présent que la détection des entités sert ensuite de point de départ à l'extraction des relations. Le rappel pourrait quant à lui être amélioré en étant plus exhaustif dans les listes d'entités constituées.

Concernant l'extraction et la validation des relations sémantiques, nous avons appliqué la méthode présentée à la Section 4.2 sur 65 Mo du corpus utilisé dans le cadre du projet Technolangue Atonant, soit environ 10 millions de mots. Les patrons d'extraction appliqués avaient été préalablement appris à partir de la totalité du corpus médical EQueR, soit environ 16 millions de mots. Contrairement au cas des entités, l'annotation manuelle de référence n'a pas été réalisée en parcourant tout le corpus mais en jugeant de la présence effective d'une des quatre relations cibles parmi les phrases abritant des relations candidates, c'est-à-dire les phrases contenant au moins deux entités compatibles avec des relations cibles. Par conséquent, seule la validation des relations candidates est évaluée ici. Pour les mesures d'évaluation, nous avons à nouveau fait appel à la précision et au rappel, définis comme suit :
– la précision représente le nombre de relations validées correctes sur le nombre total des relations validées par notre système ;
– le rappel représente le nombre de relations validées correctes par notre système sur le nombre total de relations annotées dans le corpus.
Comme dans le cas de la reconnaissance des entités, la validation des relations extraites se caractérise par une forte précision et un rappel un peu moins élevé. La différence entre précision et rappel est d'ailleurs plus accentuée dans ce cas que pour la reconnaissance des entités, un peu du fait d'une précision moyenne légèrement plus forte mais surtout par un rappel notablement moins élevé. On peut donc dire que les relations produites par la méthode que nous avons proposée sont globalement d'une bonne fiabilité mais que les patrons linguistiques appris sur le corpus médical EQueR ne couvrent pas toutes les formes par lesquelles les relations cibles

Relations	Précision	Rappel	F1-mesure
Maladie–Examen	0,92	0,63	0,74
Maladie–Médicament	0,91	0,59	0,71
Maladie–Traitement	0,92	0,69	0,78
Maladie–Symptôme	0,90	0,65	0,75
Moyenne	0,91	0,64	0,75

TAB. 2 – Résultats de la validation des relations sémantiques

se manifestent dans le corpus Atonant. La comparaison avec d'autres travaux est quant à elle difficile du fait de la diversité des types de relations considérés, des corpus et des approches adoptées. Néanmoins, il est possible de donner quelques éléments de situation. En utilisant des patrons linguistiques élaborés manuellement pour caractériser des relations d'inhibition dans des phrases extraites de Medline, (Pustejovsky *et al*., 2002) obtient ainsi une précision de 94% et un rappel de 58,9%. La Table 2 montre que nous obtenons des résultats globalement comparables en construisant ces patrons linguistiques de manière automatique. Le processus de validation des relations extraites peut également être envisagé sous l'angle de la classification : une relation candidate est alors classée comme pertinente ou non pertinente. C'est l'approche retenue par (Craven, 1999) ou par (Rosario & Hearst, 2004). En utilisant un classifieur bayésien naïf sur des relations candidates de type *subcellular–location* extraites de Medline, (Craven, 1999) fait état d'une précision de 78% et d'un rappel de 32%. Dans le cas de (Rosario & Hearst, 2004), le classifieur n'est plus seulement binaire. Il s'agit en effet de discriminer les relations intervenant entre un traitement et une maladie : 8 relations sont ainsi distinguées qui recouvrent la relation *Traite* à laquelle nous nous sommes attachés mais également des relations exprimant qu'un traitement peut prévenir une maladie ou qu'une maladie est un effet secondaire d'un traitement. (Rosario & Hearst, 2004) rapporte les évaluations menées avec plusieurs types de classifieurs et obtient les meilleurs résultats avec un réseau de neurones, la précision étant alors de 96,9%. Il est à noter que ce travail s'appuie sur des ressources plus étendues que le nôtre puisqu'il fait appel à un analyseur syntaxique de surface et qu'il exploite également la ressource sémantique que constitue le MeSH.

6 Discussion

La méthodologie proposée pour l'extraction de relations sémantiques dans le domaine médical repose sur l'identification des entités du domaine puis la validation de relations candidates extraites sur la base de la cooccurrence de ces entités en utilisant des patrons linguistiques. L'utilisation de schémas lexico-syntaxiques pour l'extraction de relations sémantiques a déjà fait l'objet de nombreux travaux. Hearst (Hearst, 1992) est l'une des premières à avoir proposé une approche fondée sur des patrons pour extraire des relations d'hyperonymie. Cependant, sa méthode, qui consiste à extraire un environnement commun à un ensemble de phrases, était essentiellement manuelle. Cette approche a été reprise et complétée par d'autres travaux, toujours dans le domaine de l'extraction de relations sémantiques, dans le but notamment d'automatiser l'extraction des patrons. La méthode développée par Ravichandran (Pantel *et al*., 2004) dont nous nous sommes inspirés se situe précisément dans cette perspective. Cette démarche s'est également avérée particulièrement productive dans des domaines de spécialité comme en

attestent par exemple les travaux rapportés dans (Finkelstein-Landau & Morin, 1999) ou (Séguéla, 1999), qui se sont focalisés sur des textes techniques.

Bien que se situant dans le droit fil de tous ces travaux, la méthode que nous avons exposée ici s'en différencie par le mode d'application des patrons linguistiques induits. Au lieu de les appliquer à la manière d'expressions régulières, nous calculons une distance entre le patron et la phrase abritant une relation candidate. Cette façon de faire autorise une plus grande souplesse dans l'application des patrons et permet également d'avoir le même mode de fonctionnement lorsque les relations sont caractérisées par des patrons, comme c'est le cas ici, et lorsqu'elles sont caractérisées par des exemples, comme dans une approche de type Memory-Based Learning. On peut même envisager ainsi de mêler les deux approches.

Une autre différence notable avec les travaux tels que (Pantel *et al*., 2004) est que les résultats de la Section 5 ont été obtenus sans utilisation d'un filtrage *a posteriori* des relations extraites. En dépit de cette absence, la précision se situe à un haut niveau sans que le rappel ne soit trop faible. Plusieurs explications complémentaires peuvent être avancées. Tout d'abord, cette extraction intervient dans un domaine spécialisé et se focalise sur des relations intervenant entre des entités spécifiques à ce domaine. Ensuite, les relations sont de type syntagmatique et non paradigmatique comme dans (Pantel *et al*., 2004). Enfin, les patrons linguistiques appris restent assez spécialisés puisqu'ils ne sont issus que de la généralisation de couples d'exemples.

7 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode d'extraction de relations sémantiques intervenant entre des entités du domaine médical. Cette méthode utilise des patrons linguistiques multi-niveaux pour valider des relations sémantiques candidates extraites des textes. Ces patrons sont appris automatiquement à partir de textes annotés en s'appuyant sur une notion de distance d'édition étendue.

La méthode proposée ici a montré des résultats encourageants en regard des travaux comparables existants. L'axe principal d'amélioration doit porter sur le rappel. Les évaluations relatives à la validation des relations ont montré que les patrons linguistiques appris ne couvrent pas toutes les manifestations des relations cibles. En outre, étant réalisées seulement à partir des phrases extraites et non de toutes les phrases du corpus d'évaluation du fait de la taille de ce dernier, elles masquent le déficit de rappel résultant de l'absence de reconnaissance des entités médicales déclenchant le processus d'extraction. Même si le niveau de reconnaissance de ces entités peut être considéré comme bon, la nécessité de reconnaître les deux entités d'une relation amplifie l'impact de leur éventuelle mauvaise reconnaissance.

Pour améliorer à la fois la couverture des patrons linguistiques et la reconnaissance des entités médicales, nous envisageons d'adopter une démarche itérative classiquement utilisée dans un tel cas : au lieu de limiter l'usage des patrons linguistiques à la seule validation des relations extraites, il est aussi possible de les utiliser pour extraire de nouvelles entités en ne fixant qu'une seule des entités d'une relation. Ces nouvelles entités viennent à leur tour enrichir la reconnaissance des entités médicales et peuvent ainsi servir à acquérir de nouveaux patrons linguistiques. Une autre voie d'amélioration du rappel est l'utilisation des ressources sémantiques existant dans le domaine médical, comme le thésaurus MeSH ou le méta-thésaurus UMLS. Il

serait ainsi possible d'inclure la vérification de relations sémantiques telles que l'hyperonymie dans la distance d'édition étendue permettant à la fois de construire les patrons linguistiques et de les appliquer. Enfin, parmi les extensions envisagées de ce travail figure également une extension de la couverture des relations de notre ontologie médicale, dont la Figure 1 ne montre qu'une partie. Nous nous sommes limités pour le moment à quatre relations mais les principes testés peuvent tout à fait être appliqués aux autres relations de cette ontologie.

Références

BESANÇON R. & CHALENDAR G. D. (2005). L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASY. In M. JARDINO, Ed., *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, p. 21–24, Dourdan : ATALA LIMSI.

CARABALLO S. A. (1999). Automatic acquisition of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 120–126.

CRAVEN M. (1999). Learning to extract relations from medline. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, Orlando, Florida, USA.

ELY J., OSHEROFF J., EBELL M., BERGUS G., LEVY B., CHAMBLISS M. & EVANS E. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ*, **319**, 358–361.

C. FELLBAUM, Ed. (1998). *WordNet : An Electronic Lexical Database*. The MIT Press.

FINKELSTEIN-LANDAU M. & MORIN E. (1999). Extracting semantic relationships between terms : Supervised vs. unsupervised methods. In *International Workshop on Ontological Engineering on the Global Information Infrastructure*, p. 71–80.

HEARST M. (1992). Automatic aquisition of hyponyms from large text corpora. In *14th International Conference on Computational Linguistics (COLING'92)*, Nantes, France.

MUKHERJEA S. & SAHAY S. (2006). Discovering biomedical relations utilizing the world wide web. In *Pacific Symposium on Biocomputing 11*, p. 164–175.

NÉDELLEC C. (2004). Machine Learning for Information Extraction in Genomics - State of the art and perspectives. In S. SIRMAKESSIS, Ed., *Text Mining and its Applications : Results of the NEMIS Launch Conference*. Springer Verlag.

PANTEL P., RAVICHANDRAN D. & HOVY E. (2004). Towards terascale knowledge acquisition. In *International Conference on Computational Linguistics (COLING'04)*, p. 771–777, Geneva, Switzerland.

PUSTEJOVSKY J., CASTANO J. & ZHANG J. (2002). Robust relational parsing over biomedical literature : Extract inhibit relations. In *PSB 2002*, p. 362–373.

ROSARIO B. & HEARST M. (2004). Classifying semantic relations in bioscience texts. In *42th Annual Conference of the Association for Computational Linguistics (ACL'04)*.

SÉGUÉLA P. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Actes de la conférence Ingénierie des Connaissances (IC'99)*, p. 79–88, Palaiseau.

VINTAR S. & BUITELAAR P. (2003). Semantic relations in concept-based cross-language medical information retrieval. In *ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM)*, Germany.