Mise en place d'une méthode de reconnaissance des signes et des symptômes dans le contexte des maladies rares

Laure Martin[1,2], Delphine Battistelli[1], Thierry Charnois[3], Marie-Christine Jaulent[4] et Laure Marelle[4]

[1] Laboratoire MoDyCo, Université de Paris Ouest Nanterre La Défense,
{laure.martin.1988,del.battistelli}@gmail.com
[2] Laboratoire GREYC, Université de Caen Basse-Normandie,
[3] Laboratoire LIPN, Université de Paris 13 Nord,
thierry.charnois@lipn.univ-paris13.fr
[4] Laboratoire LIMICS, Centre de Recherche des Cordeliers,
marie-christine.jaulent@crc.jussieu.fr
laure.lmarelle@gmail.com

Résumé : Le projet ANR Hybride[1] dans lequel s'inscrit ce travail a pour objectif la capitalisation de connaissances pour la documentation des maladies rares. En particulier, cet article s'intéresse à la détection des symptômes dans les textes médicaux, une tâche liée à la reconnaissance des entités nommées, mais cependant peu étudiée dans le cas des symptômes. Dans un premier temps, le problème spécifique de la reconnaissance des signes et symptômes est décrit d'un point de vue linguistique. Une méthodologie combinant des techniques de fouille de motifs et de traitement linguistique est ensuite proposée. Face à l'absence de corpus annoté de référence, notre approche présente l'avantage d'être faiblement supervisée. Les premiers résultats expérimentaux obtenus sont discutés ; ils ouvrent des pistes prometteuses.

Mots-clés : Traitement automatique des langues, Fouille de motifs, Reconnaissance des symptômes, Extraction d'information dans les textes médicaux, Maladies rares

1 Introduction

Le projet ANR Hybride auquel notre travail prend part a pour objectif la capitalisation de connaissances pour la documentation des maladies rares (MR). Une maladie est considérée comme rare si elle touche moins de 1 personne sur 2 000. Il existe entre 6 000 et 8 000 MR ; 30 millions de personnes en sont affectées en Europe. Les scientifiques ont besoin de rassembler et de mettre à jour régulièrement les informations concernant ces maladies, processus qui s'avère extrêmement fastidieux s'il repose sur un travail uniquement manuel. Fournir une aide à l'acquisition automatique de connaissances liées aux MR à partir de larges collections de données textuelles revêt ainsi un enjeu crucial.

Le travail décrit dans cet article s'intéresse à la détection des symptômes associés aux MR dans les résumés d'articles scientifiques. Pouvant être rapportée à la détection d'entités nommées, très étudiée dans le domaine biomédical, on constate pourtant que très peu de travaux se sont intéressés jusqu'ici à celle des symptômes.

Dans un premier temps, nous présenterons le contexte scientifique de notre étude (section 2), comprenant l'état de l'art, la présentation du scénario du projet Hybride, et le

[1] http://hybride.loria.fr/

problème de la reconnaissance des signes et symptômes en caractérisant le phénomène du point de vue linguistique. Puis, nous introduirons notre corpus et notre méthode (section 3), avant de présenter notre première expérience et ses résultats (section 4), que nous discuterons (section 5).

2 Contexte de l'étude

2.1 Scénario du projet Hybride

Orphanet est un serveur d'informations pour tous publics sur les MR et les médicaments orphelins. Parmi ses activités, Orphanet maintient une encyclopédie des MR par un processus de veille documentaire manuel. Dans l'encyclopédie, une synthèse est associée à chaque maladie. Cette synthèse est un texte informatif plus ou moins développé qui décrit l'état des connaissances sur cette maladie au moment où la synthèse a été rédigée. L'auteur de la synthèse ainsi que les sources d'information à l'origine de cette synthèse sont en général mentionnés. Les éléments de description d'une maladie sont divers : âge du début de la maladie (*onset*), prévalence, signes et symptômes, mode de transmission, causes de la maladie (étiologie), éléments du diagnostic, traitement et pronostic. L'information identifiée dans une synthèse éditée l'année X va constituer la référence pour cette maladie. Le scénario que nous mettons à l'œuvre dans le projet Hybride consiste à :

1 construire un corpus d'articles scientifiques et de documents officiels concernant cette maladie qui soient publiés postérieurement à l'année X,
2 identifier dans ce corpus les éléments de description qui se trouvent dans la synthèse,
3 évaluer le caractère novateur ou contradictoire de ces éléments par rapport à ceux qui existent déjà dans la synthèse,
4 présenter ces éléments à l'expert pour validation.

2.2 Le problème de la reconnaissance des signes et des symptômes

Nous nous intéressons ici à la fois aux symptômes et aux signes cliniques. Tous deux désignent les manifestations d'une maladie, à la différence que le symptôme (ou signe fonctionnel) décrit cette manifestation dont le patient se plaint, tandis que le signe clinique décrit cette manifestation telle qu'observée par les professionnels de santé. Par exemple, si un patient déclare avoir mal à la tête, il s'agit d'un symptôme que le médecin appellera céphalée. Il s'agit d'un signe clinique commun à plusieurs maladies et les caractères de ce signe (céphalée) ainsi que les signes qui lui sont associés permettront de le rattacher à l'une plutôt qu'à une autre.

Dans les textes, et en particulier dans les synthèses du site Orphanet ou dans les résumés d'articles scientifiques, il n'existe pas, entre signe et symptôme, de différence d'ordre morphologique ou syntaxique. La différence est purement sémantique, et le contexte linguistique ne permet pas à un non-spécialiste du domaine de faire la différence. Dans l'exemple (1), nous signalons en gras les signes cliniques et en italique les symptômes.

> (1) Cluster headache (CH) is a primary *headache* disease characterized by recurrent short-lasting attacks (15 to 180 minutes) of excruciating unilateral periorbital *pain* accompanied by **ipsilateral autonomic signs** (*lacrimation*, **nasal congestion**, **ptosis**, **miosis**, lid **edema**, and eye **redness**).

Nous n'avons pas non plus pour notre part cherché à établir une distinction. En outre, c'est bien l'ensemble des signes et des symptômes qui permet d'établir un diagnostic. Dans notre approche, nous nous intéressons donc indifféremment aux signes et symptômes.

L'observation du corpus nous a permis de constater que signes et symptômes revêtaient des formes très diversifiées. Dans leurs formes les plus simples, il s'agit de noms communs,

parfois accompagnés d'extensions du nom (2). On trouve également d'autres formes, beaucoup plus complexes, pouvant aller de la proposition à la phrase entière, c'est-à-dire qu'elles comportent un verbe et un sujet, voire un ou des compléments (3).

> (2) With disease progression patients additionally develop **weakness** and **wasting of the limb and bulbar muscles**, manifesting as **dysarthria**, **dysphonia**, **hanging jaw**, **tongue wasting**, **chewing difficulty** and **impaired mobility**.
>
> (3) Diagnosis is based on clinical presentation, and **glycemia and lactacidemia levels, after a meal (hyperglycemia and hypolactacidemia), and after three to four hour fasting (hypoglycemia and hyperlactacidemia)**.

Outre leur diversité, les unités linguistiques dénotant des signes et des symptômes présentent également un certain nombre d'ambiguïtés syntaxiques, notamment de rattachement prépositionnel et liées à la portée de la coordination : dans l'exemple (2), la première occurrence de *and* est ambigüe car on ne sait au premier abord s'il faut regrouper *weakness* et *wasting*, ou si l'on a d'un côté *weakness* et de l'autre *wasting of the limbs* (comme annoté en gras).

À ces problèmes d'ambiguïtés syntaxiques sont liées trois autres difficultés d'annotation. La première consiste à savoir délimiter correctement les unités linguistiques : l'annotation en gras dans l'exemple (4a) doit-elle comprendre les éléments en italique ? Nous avons convenu avec des experts du domaine que d'une manière générale, les informations comme les adjectifs d'intensité ou les localisations anatomiques ne faisaient pas partie des symptômes dans l'absolu ; par contre, elles peuvent être intéressantes en tant que contexte linguistique des signes et symptômes. La seconde difficulté concerne les constructions elliptiques : où nous avons deux signes distincts, nous sommes obligés de n'en annoter qu'un parce que les deux noms ont un adjectif en commun : dans l'exemple (4b), la coordination *or* n'est pas ambiguë, dans la mesure où le terme *arrest* ne fait sens qu'accompagné d'un adjectif (ici *respiratory*).

> (4) In the severe forms, **paralysis** (4a) *concerns the neck, shoulder, and proximal muscles*, followed by involvement of the muscles of the distal upper extremities, the diaphragm and respiratory muscles, which may result in **respiratory compromise or arrest** (4b).

Enfin, la dernière difficulté que nous avons rencontrée lors de l'observation du corpus est celle de l'ambiguïté entre les dénominations de signes et symptômes et de maladies. Une maladie peut être le signe d'une autre maladie. Un nom de signe ou de symptôme peut être inclus dans un nom de maladie ou inversement (5).

> (5) The adult form results in progressive limb-girdle **myopathy** beginning with the lower limbs, and affects the respiratory system, which can be the first sign of the disease.

2.3 Etat de l'art

Signes et symptômes ont été encore très peu étudiés dans le domaine de l'extraction automatique d'information dans des textes biomédicaux. Du reste, on note que ces deux types d'entités ne sont pas nécessairement désignés expressément comme tels. Ils sont par exemple englobés sous le terme « concepts cliniques » dans (Wagholikar et al., 2013), de « problèmes médicaux » dans (Uzuner et al., 2011) ou d'« information phénotypique » dans (South et al., 2009). De plus, la plupart de ces études se basent sur des corpus de rapports ou dossiers cliniques (*clinical reports / narratives*), tels que le Mayo Clinic corpus (Savova et al., 2010) ou le 2010i2b2/VA Challenge corpus (Uzuner et al., 2011). La seule exception notable est le MEDLEX Corpus (Kokkinakis, 2006), en langue suédoise, qui comporte des documents officiels, des articles scientifiques tirés de revues médicales, des documents pédagogiques, etc. Notre travail quant à lui, dans une optique de veille documentaire, se base sur un corpus de résumés d'articles scientifiques, plus faciles à obtenir de manière automatisée.

La plupart de ces systèmes utilisent des ressources lexicales, tel que les thesauri de l'Unified Medical Language System (UMLS) ou Medical Subject Headings (MeSH), pour la tâche de reconnaissance des entités nommées. UMLS regroupe plus de 100 vocabulaires contrôlés, dont MeSH, qui est un thesaurus médical générique comptant plus de 25 000 descripteurs. Cependant, comme (Albright et al., 2013) le font remarquer, UMLS n'a pas été fait dans une visée d'annotation. Ils notent ainsi que nombre de définitions de types sémantiques se recouvrent, au moins partiellement, non sans ajouter que la taille même du schéma UMLS accroît la complexité de la tâche et ralentit l'annotation, alors que seule une petite proportion des types d'annotations présents sont utilisés. C'est pourquoi ils ont préféré travailler sur les groupes sémantiques d'UMLS, plutôt que sur les types sémantiques, à l'exception des signes et symptômes, à l'origine un type du groupe sémantique « Disorders ».

Dans le contexte d'une maladie génétique, un symptôme ou un signe peuvent être phénotypiques. Le phénotype est l'ensemble des caractères observables d'une personne (morphologiques, biochimiques, physiologiques). Il résulte de l'interaction du génotype avec son milieu (l'environnement dans lequel il se développe). Comme de nombreuses maladies rares sont génétiques, de nombreux signes et symptômes de ces maladies peuvent être trouvés dans des listes d'anomalies phénotypiques. C'est pourquoi nous avons choisi de travailler avec la Human Phenotype Ontology (HPO ; Köhler et al., 2014) comme ressource lexicale première. A notre connaissance, aucune étude n'a utilisé à ce jour l'ontologie HPO. Néanmoins, il faut rappeler que les anomalies phénotypiques ne sont pas toujours des signes, et que les signes ne sont pas tous liés au phénotype. Malgré tout, nous avons décidé d'utiliser HPO, parce qu'elle comprend plus de 10 000 termes et qu'elle est facile à récupérer.

Classiquement pour ce qui relève de l'extraction d'information (et des entités nommées en particulier), ce sont des méthodes d'apprentissage qui ont été privilégiées. Plus proches de nous, nous pouvons mentionner des travaux qui évoquent la question des contextes linguistiques apparaissant autour des signes/symptômes : (Kokkinakis, 2006), après une première étape d'annotation avec MeSH, note que dans 75% des cas un symptôme apparait en cooccurrence avec d'autres symptômes dans une phrase (jusqu'à cinq) ; il peut ainsi développer de nouvelles règles d'annotation. (Savova et al., 2010) remarquent pour leur part qu'en général les signes et symptômes sont donnés en relation avec une localisation anatomique. Enfin, nous pouvons citer le système MEDLEE (Friedman, 1997), qui fournit pour chaque concept annoté son type (par exemple « problem »), sa valeur (« pain »), et ses modifieurs tels que le degré (« severe ») ou sa localisation anatomique (« chest).

Notre approche quant à elle est basée sur la combinaison des techniques du TAL et de la fouille de motifs. Nous verrons que les contextes linguistiques que nous venons d'évoquer font partie des motifs découverts automatiquement par notre outil de fouille.

3 Corpus et méthodes

3.1 Corpus et méthode générale

Le scénario mis à l'œuvre dans le projet Hybride met en relation deux types de textes : d'un côté nous avons les synthèses Orphanet à enrichir, et de l'autre les sources d'informations scientifiques fournies par les résumés d'articles scientifiques. Les deux corpus sont très différents, tant dans le fond que dans la forme. D'une part, les résumés d'articles scientifiques traitent de sujets très divers (traitements, études de cas, description d'un nouveau symptôme, etc.) et emploient un vocabulaire scientifique précis. D'autre part, nous avons les synthèses Orphanet, qui abordent tous les thèmes dans le même texte, suivant une organisation bien précise (voir section 2.1). Ces différences rendent difficile l'extraction de motifs communs aux deux types de textes, et c'est pourquoi nous avons choisi de traiter les deux corpus séparément.

Le corpus Orphanet a été constitué manuellement par un expert du domaine. Il est composé de 3 597 synthèses, rédigées entre 1996 et 2013. Les textes sont disponibles au format XML

et sont regroupés par année de mise à jour. En plus du texte de la synthèse, découpé en paragraphes[2], nous avons le titre (constitué du nom de la maladie), les différentes appellations de cette maladie, les spécialistes de la maladie qui ont relu la synthèse et la date de mise à jour.

A partir de la date de mise à jour et du titre, nous avons récupéré un corpus de résumés sur la plateforme PubMed, en veillant à ce que pour chaque synthèse, les résumés obtenus soient postérieurs à la date de dernière mise à jour. Pour chaque synthèse Orphanet, une nouvelle requête PubMed a été formulée à l'aide la date et du titre de la synthèse (correspondant au nom de la maladie). Ce processus de requêtes a pu être mis en place automatiquement à l'aide de l'outil *ebot*[3] mis à disposition par le National Center for Biotechnology Information (NCBI). Le résumé, constitué d'un seul paragraphe, est accompagné du titre et de l'identifiant PubMed de l'article. Le nombre de résumés obtenus pour chaque maladie varie de 0 à plusieurs milliers, et est amené à évoluer puisqu'il s'agit d'un corpus de veille.

Afin d'alléger le coût des prétraitements, nous n'avons travaillé qu'avec une seule liste de termes apparentés aux signes et aux symptômes, HPO. A partir de l'ontologie, nous avons constitué une liste de termes, que nous avons appliquée au corpus à l'aide d'une simple expression régulière. Même si, nous l'avons dit, l'annotation des textes avec la liste des termes d'HPO est incomplète (section 2.3), cette première annotation rapide et peu coûteuse nous permet d'amorcer le processus d'extraction, que nous allons maintenant décrire.

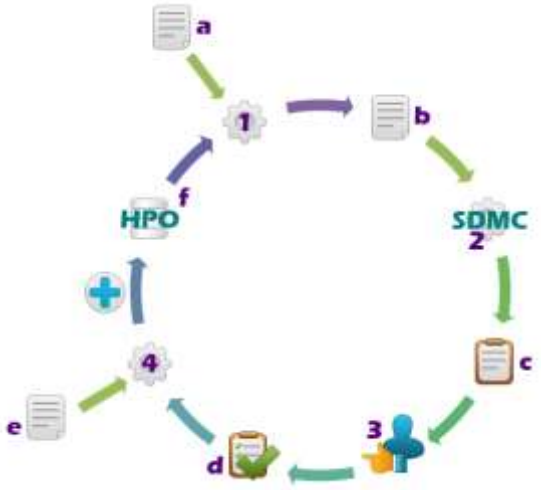

FIGURE 1 – Méthode itérative mettant en application notre méthode d'extraction des signes et symptômes. (a) premier corpus non annoté ; (b) corpus d'apprentissage annoté avec HPO ; (c) motifs obtenus avec la méthode de fouille ; (d) motifs sélectionnés par l'expert ; (e) nouveau corpus non annoté. (1) annotation simple du premier corpus avec HPO ; (2) fouille avec l'outil SDMC ; (3) sélection par un expert des motifs pertinents ; (4) application des motifs sur un corpus non annoté pour extraire de nouveaux signes et symptômes.

La figure 1 illustre les différentes étapes de l'approche. Dans l'étape 1, l'ontologie HPO (f) est utilisée pour annoter un premier corpus (a) par simple projection des termes de HPO reconnus dans les textes. Ce corpus annoté fournit un premier corpus d'apprentissage (b) pour découvrir par fouille de données (étape 2 ; pour plus de détails sur l'outil SDMC, voir ci-dessous, section 3.3) des motifs (c). Ces motifs sont ensuite validés par un expert (étape 3) en tant que patrons linguistiques (d). L'étape 4 consiste à utiliser ces patrons pour annoter de nouveaux corpus (e) et extraire de nouveaux termes de type symptôme qui viendront compléter les ressources. Le processus est ensuite réitéré (retour à l'étape 1 avec les ressources enrichies). Ce processus incrémental présente l'avantage d'être faiblement

[2] Les paragraphes sont soit d'origine, soit ont été recréés manuellement par l'expert qui a constitué le corpus.

[3] `http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi`

supervisé et non dépendant du type de corpus.

3.2 Fouille de motifs séquentiels

La découverte de motifs séquentiels a été introduite par (Agrawal et al., 1995) dans le domaine du *data mining* et adaptée par (Béchet et al., 2012) à l'extraction d'information dans les textes. Il s'agit de repérer, dans un ensemble de séquences, des enchainements d'items ayant une fréquence d'apparition supérieure à un seuil donné (dit « support »). La recherche de ces motifs s'effectue dans une base de séquences ordonnées d'items où chaque séquence correspond à une unité de texte (ici la phrase). Un item représente un mot de cette séquence (généralement la forme fléchie ou le lemme, voire la catégorie grammaticale du mot si l'on souhaite obtenir des motifs génériques). Un certain nombre de paramètres peuvent ainsi être adaptés selon l'application.

Contrairement aux approches classiques de type Machine Learning qui produisent des modèles numériques difficilement compréhensibles, la fouille de données permet la découverte de motifs symboliques et interprétables par un expert. En l'absence de corpus annotés de référence pour la reconnaissance des symptômes, une phase de validation manuelle des motifs est nécessaire, et souvent, le nombre de motifs extraits reste important. Pour pallier ce problème, (Béchet et al., 2012) proposent l'ajout de contraintes pour diminuer la quantité de motifs retournés. Dans la lignée de ces travaux, nous utilisons l'outil d'extraction de motifs séquentiels SDMC[4], lequel permet l'utilisation de différentes contraintes ainsi que l'extraction de représentation condensée (motifs sans redondance).

Nous avons adapté la fouille de motifs à notre domaine d'application. Ainsi, notre approche propose tout d'abord de prétraiter un corpus grâce l'outil TreeTagger (Schmid, 1994), après un découpage en phrases correspondant à des séquences, afin d'obtenir différents types d'items : forme fléchie, lemme, catégorie grammaticale. Pour limiter le nombre de motifs retournés par l'outil, nous introduisons un ensemble de contraintes spécifiques à notre application : des contraintes linguistiques d'appartenance (nous pouvons par exemple choisir de ne retourner que des motifs contenant au moins un nom de symptôme) mais aussi une contrainte dite de "gap" (Dong & Pei, 2007), correspondant aux trous possibles entre items dans le motif. Ainsi un gap d'une valeur maximale *n* signifie qu'au maximum *n* items (mots) sont présents entre chaque item du motif dans les séquences (phrases) correspondantes.

4 Expérimentation et premiers résultats

Après avoir annoté un premier corpus de 306 717 résumés PubMed avec la liste des anomalies phénotypiques proposée par l'ontologie HPO, nous avons sélectionné 10 000 phrases comportant 13 477 unités annotées. Ces signes et symptômes ont été remplacés par le mot-clef SYMPTOM dans le corpus, afin de faciliter la découverte de motifs. Puis nous avons utilisé l'outil SDMC (voir section 3.3) afin d'extraire les motifs du corpus avec les contraintes suivantes : nous avons cherché les motifs maximaux d'un support minimal de 10, d'une longueur comprise entre 3 et 50 mots et avec une contrainte de gap maximale et minimale fixée à 0. Nous avons effectué la fouille sur les lemmes.

Le résultat de la fouille nous a fourni 988 motifs, dont 326 contenant au moins une fois le mot-clef « symptom » utilisé pour remplacer les termes annotés à l'aide de la liste HPO. Nous pouvons d'ores et déjà faire quelques remarques sur les motifs ainsi obtenus :

- plusieurs symptômes sont associés à un troisième terme non annoté par HPO mais qui pourrait bien avoir le statut de symptôme : {symptom}{symptom}{and}{stress} ;
- les limites de l'annotation avec HPO (section 2.3) sont rendues visibles par le contexte : {disease}{such}{as}{symptom} ;

[4] Disponible en ligne : `http://sdmc.greyc.fr/`

- certains motifs récurrents dans les synthèses Orphanet sont également présents, comme {be}{associate}{with}{symptom} ou {characterize}{by}{symptom} ;
- quelques contextes temporels et d'ordonnancement sont présents, comme {@card@}{%}{follow}{by}{symptom};
- le terme « patient » est très présent ({patient}{have}{severe}{symptom}), mais après évaluation, il s'avère que la plupart de ces motifs s'appliquent à des maladies, plus qu'à des signes ou des symptômes ;
- autre contexte assez présent aussi, celui des parties du corps, qui viennent préciser la localisation du signe en question : {frontotemporal}{symptom}{ftd}.

Dans un premier temps, un expert linguiste a sélectionné manuellement les motifs qui paraissaient les plus intéressants ; ces motifs ont été classés dans trois catégories, selon qu'ils semblaient impliquer la présence de signes ou de symptômes fortement (43 motifs), moyennement (309 motifs) ou faiblement (45 motifs). Dans un deuxième temps, ces motifs ont été appliqués sur un nouveau corpus de résumés d'articles scientifiques, afin d'annoter les contextes des signes et des symptômes. A l'heure actuelle, nous n'avons appliqué que les motifs « forts ».

25 résumés ont été sélectionnés au hasard parmi tous les articles publiés au cours du dernier mois et traitant de la maladie de Pompe. Ces 25 articles ont été annotés manuellement par un expert, constituant ainsi un *gold standard*. Ensuite, nous avons confronté nos contextes annotés automatiquement à ces annotations manuelles. Si la phrase annotée manuellement comportait effectivement un signe, nous avons considéré l'annotation du contexte comme pertinente. Parmi les 25 résumés (255 phrases), notre méthode a permis l'annotation de 27 contextes. 23 étaient corrects, 4 étaient faux ; 70 phrases ont été omises. Ainsi, notre système a obtenu, pour cette première approche, un rappel de 23,7% et une précision de 82,2% (soit une F-mesure de 36,8%). Sur les 23 annotations correctes, 7 dénotent de signes ou symptômes qui ne sont pas mentionnés dans la synthèse Orphanet, ce qui est un premier résultat prometteur au regard de notre objectif de veille (voir le scénario évoqué en section 2.1).

5 Discussion

L'ambiguïté entre signes/symptômes et maladies est cause de 3 des 4 erreurs relevées lors de notre évaluation. La quatrième erreur désigne en réalité un test diagnostique ; cette erreur démontre que causes et conséquences (effets secondaires) des maladies sont difficiles à distinguer pour des non-spécialistes. La plupart des phrases omises par notre système comportaient des signes et des symptômes exprimés par des unités linguistiques complexes (comportant un verbe, un sujet et des compléments), comme « Levels of creatinkinase in serum were high » (36% des phrases omises sont concernées) ; on y trouve en particulier des phrases entières, des propositions coordonnées et quelques propositions subordonnées relatives. La difficulté d'annotation de ces unités linguistiques complexes est due au fait qu'elles n'ont pas de contexte à proprement parler, dans la mesure où tous les éléments font partie de l'unité à annoter. Nous ne sommes plus ici en présence d'entités nommées, et nous devons envisager de nouvelles méthodes d'annotation pour ces signes et symptômes. 27% de ces phrases ont pour sujet des mutations génétiques, qui peuvent être considérées comme des causes de la maladie ou comme des signes cliniques selon le contexte. Les autres phrases contenaient des motifs qui n'avaient pas été sélectionnés par l'expert linguiste, mais qui peuvent être facilement ajoutés.

L'annotation des contextes de signes et des symptômes n'est qu'une étape vers l'annotation des signes et des symptômes eux-mêmes. Nous avons actuellement deux hypothèses sur la délimitation des signes et des symptômes à partir d'un seul contexte. Nous pensons qu'une analyse syntaxique nous permettrait de délimiter certaines structures syntaxiques récurrentes, comme les signes compléments d'agent de certains verbes ou les énumérations. Le découpage en *chunks* est également une piste intéressante, qui peut

permettre d'obtenir les unités linguistiques correspondant aux signes ou des symptômes.

En outre, la sélection des motifs se fait manuellement à l'heure actuelle. L'idée est de mettre en place certaines règles symboliques qui permettront de faire une présélection automatique des motifs afin de faciliter la tâche des experts et d'accélérer le processus.

Nous avons également l'intention de comparer nos résultats avec ceux produits par des CRFs. Tout d'abord, les paramètres seront classiques (nous utiliserons, entre autres, les sacs de mots), puis nous ajouterons les contextes obtenus avec la fouille de texte aux paramètres. Cela nous permettra de comparer notre méthode aux autres travaux.

Enfin, nous allons développer une interface d'évaluation, afin de faciliter le travail de l'expert. En l'absence de corpus comparables, l'évaluation ne peut être que manuelle. Notre échantillon actuel de 25 résumés n'est qu'un début, et il est nécessaire de l'agrandir afin de renforcer l'évaluation.

Remerciements

Ce travail bénéficie du soutien du projet Hybride ANR-11-BS02-002.

Références

AGRAWAL R. et SRIKANT R. (1995). Mining sequential patterns. In *Proceedings of ICDE'95*.

ALBRIGHT D., LANFRANCHI A., FREDRIKSEN A., STYLER W., WARNER C., HWANG J., CHOI J., DLIGACH D., NIELSEN R., MARTIN J., WARD W., PALMER M. & SAVOVA G. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. In *Journal of the American Medical Informatics Association* vol.20, p.922-930

BECHET N., CELLIER P., CHARNOIS T., CREMILLEUX B. (2012). Discovering linguistic patterns using sequence mining. In *Proceedings of Springer LNCS, 13th International Conference on Intelligent Text Processing and Computational Linguistics - CICLing'2012* vol.1, p.154-165

DONG G. and PEI J. (2007). *Sequence Data Mining*. Springer

FRIEDMAN C. (1997). Towards a Comprehensive Medical Language Processing System: Methods and Issues. In *Proceedings of the AMIA Annual Fall Symposium*, p.595-599

KÖHLER S., DOELKEN S. & 40 auteurs. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. In *Nucleic Acid Research* vol.42, p.966-974

KOKKINAKIS D. (2006). Developing Resources for Swedish Bio-Medical Text Mining. In *Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine (SMBM)*

SAVOVA G., MASANZ J., OGREN P., ZHENG J., SOHN S., KIPPER-SCHULER K. & CHUTE C. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. In *Journal of the American Medical Informatics Association* vol.17, p.507-513

SCHMID H. (1994) Probabilistic Part-of-Speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*

SOCHER R., BAUER J., MANNING C. & NG A. (2013). Parsing With Compositional Vector Grammars. In *Proceedings of ACL 2013*

SOUTH B., SHEN S., JONES M., GARVIN J., SAMORE M., CHAPMAN W. & GUNDLAPALLI A. (2009). Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. In *Summit on Translational Bioinformatics 2009*, p.1-32

UZUNER Ö., SOUTH B., SHEN S. & DUVALL S. (2011). 2010 i2b2/VA Challenge on concepts, assertions, and relations in clinical text. In *Journal of the American Medical Informatics Association* vol.18, p.552-556

WAGHOLIKAR K., TORII M., JONNALAGADDA S. & LIU H. (2013). Pooling annotated corpora for clinical concept extraction. In *Journal of Biomedical Semantics* 2013, 4:3