Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux

Mike Donald Tapi Nzali, Sandra Bringay, Christian Lavergne, Thomas Opitz, Jérôme Azé, Caroline Mollevi

Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux

Mike Donald Tapi Nzali[1,2], Sandra Bringay[2,3], Christian Lavergne[1,3], Thomas Opitz[4], Jérôme Azé[2], Caroline Mollevi[5]

[1] I3M, Université Montpellier, France
mike-donald.tapi-nzali@univ-montp2.fr, christian.lavergne@univ-montp2.fr

[2] LIRMM, Université Montpellier, France
sandra.bringay@lirmm.fr, jerome.aze@lirmm.fr

[3] Université Paul Valery Montpellier, France

[4] Biostatistique et Processus Spatiaux (BioSP), INRA Avignon, France
thomas.opitz@paca.inra.fr

[5] Unité de biostatistique, Institut de Cancérologie de Montpellier, France
Caroline.Mollevi@icm.unicancer.fr

Résumé : De nos jours, les médias sociaux sont de plus en plus utilisés par les patients et les professionnels de santé. Les patients, généralement profanes dans le domaine médical, utilisent de l'argot, des abréviations et un vocabulaire qui leur est propre lors de leurs échanges. Pour analyser automatiquement les textes des réseaux sociaux, l'acquisition de ce vocabulaire spécifique est nécessaire. En nous appuyant sur un corpus de documents issus de messages de médias sociaux de type forums et Facebook, nous décrivons la construction d'une ressource lexicale qui aligne le vocabulaire des patients à celui des professionnels de santé. Ce travail permettra, d'une part d'améliorer la recherche d'informations dans les forums de santé et d'autre part, de faciliter l'élaboration d'études statistiques basées sur les informations extraites de ces forums.

Mots-clés : Extraction d'information, Médias sociaux, Vocabulaire patient

1 Introduction

Les vocabulaires contrôlés (e.g. SNOMED, MeSH, UMLS, etc.) jouent un rôle clé dans les applications biomédicales de fouille de textes. Ces vocabulaires contiennent seulement les termes utilisés par les professionnels de santé. Depuis 10 ans, des vocabulaires dédiés aux consommateurs de soins de santé (Consumer Health Vocabularies - CHV), ont également été crées (Zeng & Tse, 2006). Ces CHV lient des mots de tous les jours se rapportant au domaine de la santé à des mots d'argot technique utilisés par les professionnels de santé.

Dans cet article, nous proposons une méthode semi-automatique pour construire un tel CHV pour la langue française. Par exemple, nous cherchons à relier le mot "onco" utilisé par les patients à "oncologue" utilisé par les professionnels de santé. L'originalité de notre approche est d'utiliser les textes rédigés par les patients (PAT Patient-Authored Text), provenant des messages issus des médias sociaux de type forums ou Facebook, ainsi que la structure de l'encyclopédie universelle collaborative Wikipédia. Notre méthode a été expérimentée avec succès sur un jeu de données réelles dans le domaine du cancer du sein. Elle a été validée automatiquement en utilisant la ressource collaborative du site JeuxDeMots.org. Une validation manuelle a été également réalisée par 4 personnes, dont un expert du domaine du cancer du sein.

Cet article est organisé comme suit. Dans la section 2, nous motivons notre travail et donnons un état de l'art rapide. Dans la section 3, nous décrivons chaque étape de la méthode. Dans la

section 4, nous présentons le cadre expérimental utilisé pour évaluer les performances de cette méthode. Dans la section 5, nous discutons des premiers résultats. Finalement, dans la section 6, nous concluons et donnons quelques perspectives à ces travaux.

2 Motivations et état de l'art

Selon une enquête réalisée en 2011 par la fondation HON [1], Internet est devenu la deuxième source d'information des patients après les consultations chez les médecins. 24% de la population utilise Internet pour trouver des informations sur leur santé au moins une fois par jour (et jusqu'à 6 fois par jour) et 25% au moins plusieurs fois par semaine. Ces « patients 2.0 » sont motivés par un accès facile à Internet à domicile, le manque général de temps pour des consultations plus classiques, un soutien humain (surtout pour les maladies chroniques), la nécessité de connaître les expériences des autres, ainsi que le désir d'obtenir plus d'informations avant ou après une consultation (Hancock *et al.*, 2007; Merolli *et al.*, 2013). En maintenant l'anonymat, ces médias sociaux (forums, groupes Facebook) leur permettent de discuter librement avec d'autres utilisateurs, usagers, personnes, et aussi avec des professionnels de santé. Ils parlent de leurs résultats médicaux et de leurs options de traitement, mais ils reçoivent également un soutien moral.

Dans des travaux précédents (Opitz *et al.*, 2014), nous nous sommes intéressés à l'étude de la qualité de vie des patientes atteintes d'un cancer du sein à partir des médias sociaux. Nous avons cherché à capturer et quantifier ce que les patientes expriment dans les forums à propos de leur qualité de vie. Une importante limitation à ces travaux vient du type de textes traités. En effet, la plupart des patients sont des profanes dans le domaine médical. Lors de leurs échanges, ils utilisent des mots d'argot, des abréviations et un vocabulaire spécifique construit par la communauté en ligne, à la place des termes médicaux que l'on retrouve dans les ressources terminologiques utilisées par les professionnels de santé comme la SNOMED (Nomenclature systématisée de médecine) [2], le MeSH (Medical Subject Headings) [3], l'UMLS (Unified Medical Language System) [4]. Les méthodes de fouille de textes mises en œuvre ont montré leurs limites à cause de ce vocabulaire particulier. Nous nous proposons donc dans cet article de construire un vocabulaire dédié aux « consommateurs de soins de santé » (Consumer Health Vocabularies - CHV).

Initialement, la création de ces CHV a été motivée par la réduction des écarts de connaissances entre les patients et les professionnels de santé (Zeng *et al.*, 2007). En effet, la littérature montre que la compréhension par les patients de la terminologie médicale est essentielle pour appréhender leur maladie et pour participer au processus de décision médicale. En outre, les communications réussies patient-médecin sont intrinsèquement liées à la confiance que le patient a envers son médecin (Fiscella *et al.*, 2004). S'il ne comprend pas de quoi le médecin lui parle, le patient est moins enclin à lui faire confiance. Certains chercheurs ont ainsi utilisé des CHV pour améliorer la lisibilité des documents médicaux (Wu *et al.*, 2013) ou du dossier patient électronique (Ramesh *et al.*, 2013) par les non-experts. (Doing-Harris & Zeng-Treitler,

1. HON (Health On the Net) How Do General Public Search Online Health Information ? Avril 2011
2. `http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html`
3. `http://mesh.inserm.fr/mesh/`
4. `http://www.nlm.nih.gov/research/umls/`

2011) ont proposé une méthode pour générer automatiquement des termes candidats à traiter par des humains pour inclusion dans un CHV. Ils n'apparient pas automatiquement les termes des patients à ceux des médecins comme nous allons le proposer dans cet article. Actuellement, seuls deux CHV sont disponibles : 1) MedlinePlus [5], librement disponible, est produit par la National Library of Medicine ; 2) Open and collaborative Consumer Health Vocabulary (CAO CHV) [6] est inclus dans l'UMLS. À notre connaissance, en français, il n'existe pas de CHV.

Dans les médias sociaux (forums, Facebook), le volume des textes rédigés par les patients (PAT Patient-Authored Text) est de plus en plus important (MacLean & Heer, 2013). Si de tels PAT ne sont pas suffisamment précis pour des objectifs scientifiques, ils donnent accès à de très nombreuses descriptions de l'expérience des patients, sur un large éventail de sujets, en temps réel. Au cours des cinq dernières années, il y a eu un intérêt croissant dans l'exploitation de ces PAT comme outil pour la santé publique, par exemple pour des analyses de la propagation de la grippe (Sadilek *et al.*, 2012) ou la découverte d'effets secondaires grâce à des sites comme CureTogether [7] et PatientsLikeMe [8]. Dans cet article, notre objectif est d'utiliser les PAT issus des médias sociaux en entrée d'une méthode semi-automatique permettant de construire un CHV français pour le domaine du cancer du sein, en recueillant différents types d'expressions de patients, comme des abréviations, des fautes d'orthographe fréquentes ou des mots de tous les jours détournés par les non experts pour parler de leurs maladies.

L'originalité de notre approche est d'utiliser l'architecture de l'encyclopédie universelle collaborative Wikipédia [9] pour rapprocher des termes utilisés par les patients et des termes utilisés par des professionnels de la santé. Wikipédia est une encyclopédie sur le Web multilingue qui couvre de très nombreux domaines. La version française, en date du 25 février 2015 contient plus d'un million et demi d'articles. Les articles étant finement structurés, Wikipedia a été utilisée avec succès dans des applications de questions/réponses (Buscaldi & Rosso, 2006), de catégorisation de textes (Wang *et al.*, 2009). Plus particulièrement, on trouve des approches permettant de calculer la parenté sémantique entre des termes (Ponzetto & Strube, 2006; Gabrilovich & Markovitch, 2007). Ces derniers ont développé une technique permettant de représenter le sens des mots dans un espace de dimension élevée de concepts issus de Wikipedia. (Chernov *et al.*, 2006) ont utilisé les liens entre les catégories présentes sur Wikipédia pour extraire de l'information sémantique. (Witten & Milne, 2008) utilisent plutôt les liens entre les articles de Wikipedia pour déterminer la proximité sémantique entre les mots. Dans ce travail, nous allons comme (Witten & Milne, 2008), utiliser la structure de liens entre les termes Wikipédia pour rapprocher le vocabulaire des patients, de celui des médecins.

3 Méthodes

La figure 1 illustre la méthode proposée, structurée en 5 étapes. Cette méthode prend en entrée une ressource médicale à laquelle nous allons apparier les termes des patients. Nous avons

5. http://www.nlm.nih.gov/medlineplus/
6. http://www.consumerhealthvocab.org/
7. http://curetogether.com/
8. http://www.patientslikeme.com/
9. http://fr.wikipedia.org/wiki/Wikipédia:Accueil_principal

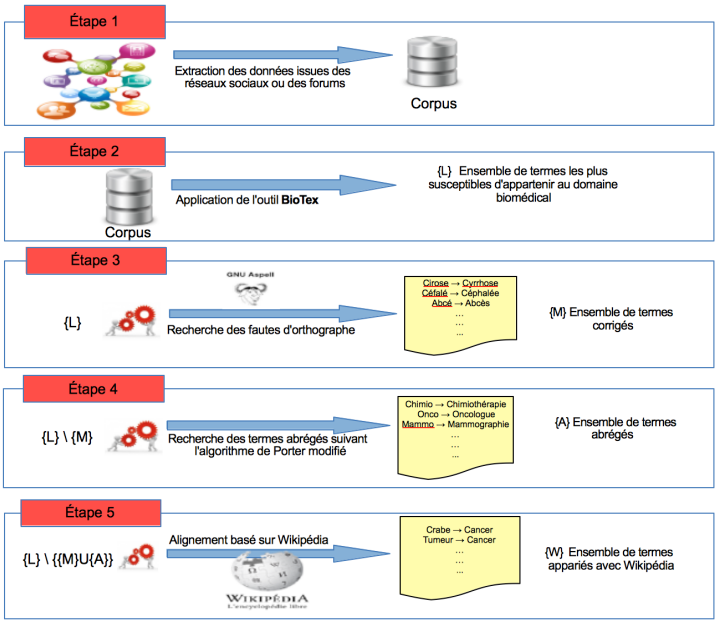

FIGURE 1 – Diagramme d'activité utilisé dans notre algorithme.

choisi comme ressource de référence le vocabulaire donné sur le site de l'INCa [10] composé de 1 227 termes, tous présents dans le MeSH en version française, que nous noterons INCa.

Étape 1 : Développement du corpus de messages. Nous utilisons des messages issus du réseau social Facebook et de forums échangeant sur le cancer du sein. Les groupes de paroles Facebook facilitent la connexion avec d'autres patientes ou associations de patientes. Les groupes permettent de publier des mises à jour, des photos ou des documents et d'envoyer des messages à tous les membres du groupe. La figure 2 correspond à un post commenté par 5 membres. Dans le post initial, apparaît l'abréviation *chimio*. Dans la première réponse, apparaît la faute d'orthographe *catheter* pour *cathéter*. Dans la troisième réponse, on trouve le terme *rdv* pour *rendez-vous*. Nous avons récolté ainsi 96 792 messages publiés par 1 389 membres entre 2010 et 2014 des groupes Facebook publics tels que *Cancer du sein, Octobre rose 2014, Cancer du sein - breast cancer, brustkrebs*. Nous avons aussi travaillé avec les données provenant du forum *lesimpatientes.com*, nous avons récolté 134 334 messages provenant de 4 627 utilisateurs. Chaque document du corpus contient l'ensemble des messages d'un utilisateur d'un forum de santé ou d'un groupe Facebook. Dans ce travail, nous travaillons uniquement sur les textes et n'utilisons aucune autre métadonnée. Un avantage de notre approche est qu'aucun traitement spécifique n'a pas été effectué sur ces messages (pas de correction automatique, ni de lemmatisation).

Étape 2 : Extraction des termes candidats à partir du corpus. À partir du corpus, nous cherchons les termes ayant une grande probabilité d'appartenir au domaine médical. Pour cela, nous utilisons l'outil BioTex (Lossio-Ventura *et al.*, 2014a). BioTex est une application d'extraction automatique de termes biomédicaux qui met à disposition un ensemble de mesures sta-

10. http://www.e-cancer.fr/cancerinfo/ressources-utiles/dictionnaire/

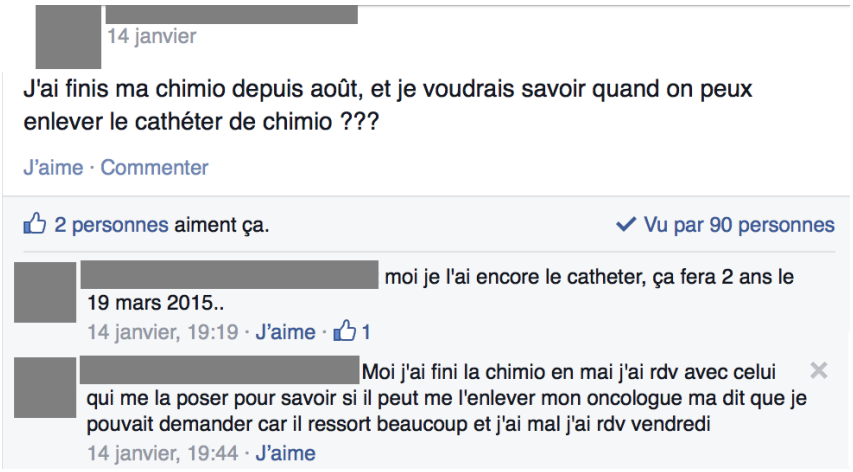

FIGURE 2 – Posts anonymisés et commentés par des utilisateurs d'un groupe Facebook.

tistiques pour la sélection de ces termes. La sélection est essentiellement basée sur la fréquence d'apparition et la construction linguistique qui doit être similaire à celle des termes présents dans les ressources médicales de type MeSH. Pour cela, 200 motifs linguistiques ont été utilisés (voir table 1). La mesure choisie est *LIDF-value (Linguisitic patterns, IDF, and C-value information)* Lossio-Ventura *et al.* (2014b) car (Lossio-Ventura *et al.*, 2014c) ont démontré que cette mesure donne de meilleurs résultats comparés à d'autres comme *TF-IDF, Okapi, C-value*. À l'issue de cette étape, nous obtenons en sortie un ensemble $T = t_1, ..., t_N$ de N **n**-grammes (n $\in [1..4]$), dont certains ne sont pas répertoriés dans l'INCa, que nous allons utiliser dans les étapes 2, 3 et 4 décrites ci-dessous. Il est important de noter que nous obtenons ici des candidats composés de plusieurs mots. Ces candidats sont spécifiques aux textes des patients traitant des sujets médicaux.

Motif	**Texte instantiant le motif**
Nom Adj	Echographie mammaire
Nom Prep :det Nom	Cancer du sein
Nom Prep NomPropre	Maladie d'Alzheimer

TABLE 1 – Exemples de motifs linguistiques utilisés dans BioTex

Étape 3 : Correction orthographique des termes candidats mal orthographiés. À partir des mots identifiés à l'étape 2, fréquemment utilisés par les patients, on recherche ceux qui correspondent à des fautes d'orthographes courantes. Nous cherchons à apparier tous les termes $t_i \in T$, avec un mot bien orthographié présent dans l'INCa. Pour cela nous utilisons le logiciel Aspell [11] pour obtenir un ensemble $M = \{m_1, m_2, ..., m_m\}$ de m propositions de corrections du mot t_i et ne conservons que les propositions présentes dans l'INCa. Nous utilisons ensuite la

11. `http://aspell.net/`

mesure de Levenshtein pour calculer la distance entre le terme t_i et chaque terme m_j. La mesure de Levenshtein entre deux termes est le nombre minimum de modifications à caractère unique nécessaires pour changer t_i en m_j. Seul les termes dont la distance est inférieure ou égale à 2 ne sont conservés comme appariement. Trois autres conditions sont également nécessaires : 1) les mots appariés doivent commencer par la même lettre ; 2) la longueur des mots appariés est de plus de trois caractères ; 3) la comparaison est insensible à la casse. Si toutes les conditions sont vérifiées, le terme t_i est associé au terme m_j avec un $poids(m_j, t_i) = 1/|M|$. La table 2 présente quelques fautes d'orthographe fréquemment rencontrées.

Termes biomédicaux	**Termes patients**
cirrhose	Cyrose
abcès	abcé
métastase	metastase

TABLE 2 – Équivalent entre termes biomédicaux et termes patients (contenant des fautes d'orthographe)

Étape 4 : Recherche des termes abrégés. La plupart des expressions biomédicales sont longues (composées de 2, 3 mots voir plus). Très souvent, ces expressions sont tronquées par les patients. À partir des mots identifiés à l'étape 2, fréquemment utilisés par les patients, on recherche ceux qui correspondent à des abréviations. Pour cela, nous avons adapté l'algorithme de (Paternostre *et al.*, 2002) en utilisant la liste des suffixes les plus utilisés dans le domaine biomédical (e.g : logie, logue, thérapie, thérapeute...). Pour un terme $t_i \in T$, on obtient un ensemble $A = \{a_1, a_2, ..., a_k\}$ de k propositions d'abréviations incluses dans l'INCa. Le terme t_i est associé à une abréviation a_j avec un $poids(a_j, t_i) = 1/|A|$. Des exemples de termes appariés avec cette méthode sont listés dans la table 3.

Termes biomédicaux	**Termes patients**
Oncologue	Onco
Chimiothérapie	Chimio
mammographie	mammo

TABLE 3 – Équivalent entre termes biomédicaux et termes patients (abréviations)

Étape 5 : Alignements basés sur Wikipédia. Nous nous intéressons ici à tous les termes produits à l'étape 2 qui ne contiennent ni des mots comportant des fautes d'orthographe fréquentes (repérées à l'étape 3), ni des abréviations (repérées à l'étape 4). Pour cela nous travaillons sur l'architecture de la ressource encyclopédique Wikipédia que nous interrogeons grâce à son API [12]. Dans cette encyclopédie un terme (*mot*) référencé est décrit par une page (*http ://fr.wikipedia.org/wiki/**mot***) et est lié à d'autres termes eux mêmes décrits par d'autres pages. Les pages (mots) liées à un terme se retrouvent dans une page dédiée

12. `http://fr.wikipedia.org/w/api.php?`

(*http ://fr.wikipedia.org/wiki/Spécial :Pages_liées/**mot***). Certaines relations entre termes Wikipédia sont typées (e.g. synonymie). Sur la partie droite de la figure 3, on retrouve la page du terme *Tumeur* et sur la partie gauche, les termes liés. Soit *W* l'ensemble des termes liés par Wikipédia à un terme t_i et appartenant à la ressource INCa. Un terme t_i est associé à un terme w_i selon un poids calculé avec la formule 2. Ce poids est important pour éliminer des associations comme *tumeur* et *jules cesar* dans l'exemple de la figure 3 car la page tumeur ne contient pas de référence à Jules César alors que cette dernière en contient (référence à son état de santé).

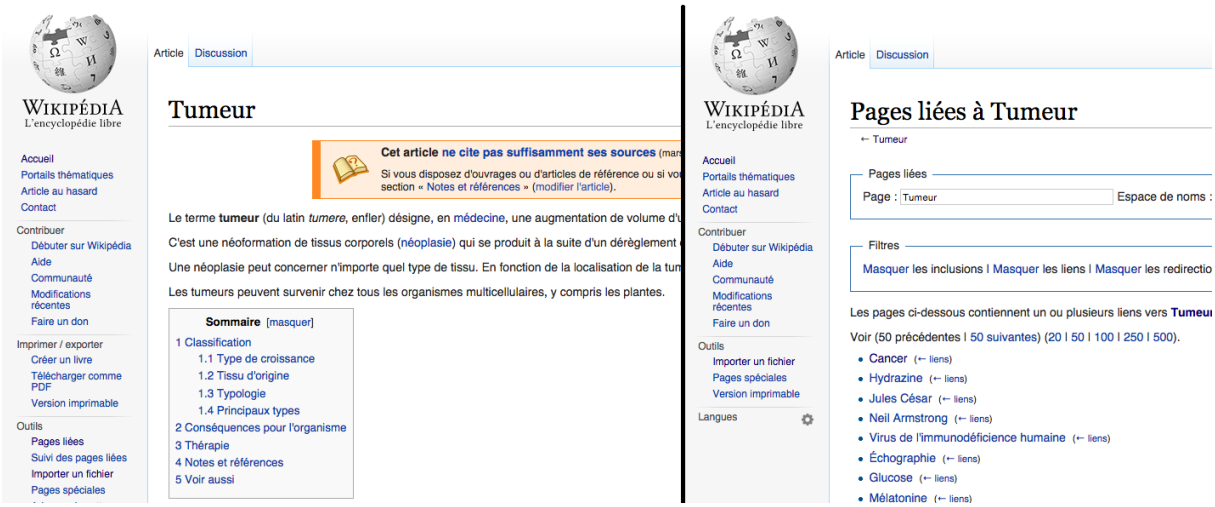

FIGURE 3 – Page Wikipédia et page liée.

$$MoyNb(w_k, t_n) = \frac{Nb(w_k, PageW(t_n)) + Nb(t_n, PageW(w_k))}{2} \quad (1)$$

$$Poids(w_j, t_i) = \frac{MoyNb(w_j, t_i)}{\sum_{k=1}^{|W|} MoyNb(w_k, t_i)} \quad (2)$$

où $Nb(a, PageW(b))$ est la fréquence d'apparition du terme a dans la page wikipédia du terme b.

4 Résultats

À l'issu du processus précédent, nous avons obtenu l'ensemble de K relations r_i avec $i \in [1, K]$. Chaque relation r_i relie un mot patient pat_i [13], avec un mot médecin bio_i [14]. Chaque relation est associée à une méthode d'obtention $meth \in$ *{orthographe,abréviation,association}* et à un poids $poids \in [0, 1]$. Dans cette section, nous présentons les deux méthodes de validation utilisées (automatique et manuelle) et les différents résultats obtenus. Comme les associations détectées automatiquement dépendent beaucoup du contenu de Wikipédia, la validation finale

13. Les pat_i sont les termes issus du corpus
14. Les bio_i sont les termes du dictionnaire fourni par l'INCa

manuelle est importante pour présentir les faiblesses des associations obtenues avec les méthodes quantitatives.

4.1 Validation automatique

Nous validons automatiquement des relations r_i, si l'un des deux critères suivants est vérifié :

— Le poids de la relation est égale à 1. Par exemple, pour une faute d'orthographe avec une seule possibilité de correction, nous considérons la correction validée.

— La paire pat_i - bio_i existe dans le dictionnaire de relations fournis par le jeux contributif www.JeuxDeMots.org, dont le but est de construire un vaste réseau lexical-sémantique (Lafourcade & Joubert, 2012). Cette ressource, construite par les internautes, rassemble 112 types de relations dont 179 578 occurences de la relation synonymie. L'avantage de cette validation est que nous obtenons une étiquette supplémentaire pour typer les relations.

Des exemples de relations validées automatiquement sont présentées dans le tableau 4. Sur les 432 relations obtenues après exécution de notre programme, 211 relations ont été validées automatiquement, soit 49% des relations. Nous avons donc 169 relations de type « association », 32 relations de type « erreur », et 10 relations de type « abréviation ».

Terme Patient	**Terme biomédical**	**Relation**
Chir	Chirurgie	Abréviation
Chimio	Chimiothérapie	Abréviation
mammo	mammographie	Abréviation
hopital	hôpital	Erreur Orthographique
cheuveux	cheveux	Erreur orthographique
radiotherapie	radiothérapie	Erreur orthographique
♡	Cœur	Association
tumeur	cancer	Association
chute des cheveux	Alopécie	Association

TABLE 4 – Exemples de termes validés automatiquement en utilisant JeuxDeMots

4.2 Validation manuelle

Toutes les relations r_i n'ayant pas pu être validées automatiquement sont présentées à l'expert pour validation manuelle [15]. Le tableau 5 présente des exemples de résultats validés manuellement. Les annotations ont été faites par 4 personnes, dont un expert du domaine du cancer du sein.

Lors de la validation manuelle, trois choix sont proposés à l'utilisateur : 1) **Yes** : pour valider la relation ; 2) **No** : pour invalider la relation ; 3) **Neutre** : l'annotateur ne sait pas. Nous considérons qu'une relation r_i est validée si le nombre de « Yes » est supérieur au nombre de « No » et de « Neutre ».

15. Un image de l'interface de valisation est présente à cette url : `http://www.lirmm.fr/~tapinzali/Validation/Validation.php`

Suite à la validation automatique de 214 relations, il nous restait 218 relations à valider manuellement. Après consensus entre les différents annotateurs, 93 relations sur les 218 ont été validées. Un coefficient de kappa de Fleiss a été calculé pour mesurer l'accord inter-annotateur, nous obtenons un ***k*** de **0,21** (accord faible, dû à la variabilité individuelle du jugement des annotateurs sur l'intérêt médical des termes).

Terme Patient	**Terme biomédical**	**Relation**
Psy	Psychologue	Abréviation
Onco	Oncologue	Abréviation
gynéco	gynécologue	Abréviation
constipation	Laxatif	Association
libido	Sexologie	Association
morphine	Douleur	Association

TABLE 5 – Exemples de termes validés manuellement

5 Discussions

Finalement, sur les deux corpus (10 Mo chacun), nous avons extrait plus de 432 relations. Ce nombre de relations augmente lorsque nous diminuons les seuils sur LIDF-value à l'étape 2 décrite dans la section 3 par exemple. Toutefois, le temps d'exécution devient alors très grand (45 minutes pour un corpus de 1Mo), car diminuer le seuil implique de prendre en compte de nouveaux termes à traiter. Par ailleurs, le nombre limité de relations obtenues s'explique également par le nombre de termes médecin cibles auxquels nous cherchons à apparier les termes des patients (ceux de l'INCa qui contient uniquement 1 227 termes). En effet, nous cherchons à créer une ressource pour le cancer du sein et non une ressource généraliste.

Comme (Doing-Harris & Zeng-Treitler, 2011), nous avons cherché à évaluer la précision globale de notre méthode. Pour cela nous avons utilisé la formule 3.

$$P = \frac{|R_a| + |R_m|}{|R|} \tag{3}$$

où R_a est l'ensemble contenant la liste des relations validées automatiquement, R_m est l'ensemble contenant la liste des relations validées manuellement et R est l'ensemble des relations ayant été fournies en sortie par notre outil. Nous avons obtenu une précision ***P*** de 71% et validé 307 relations sur les 432 obtenues. La figure 4 présente les résultats obtenus selon le type des relations, sur ce premier jeu de données nous montrent que sur des données textuelles bruitées biomédicales extraites extraites des forums de santé et des réseaux Facebook, nous obtenons de meilleurs résultats que ceux de (Doing-Harris & Zeng-Treitler, 2011). Ces auteurs ont créé un CHV en langue anglaise. Sur 88 994 n-grammes, ils ne trouvent que 774 relations et n'en valident que 237, soit 31% de précision.

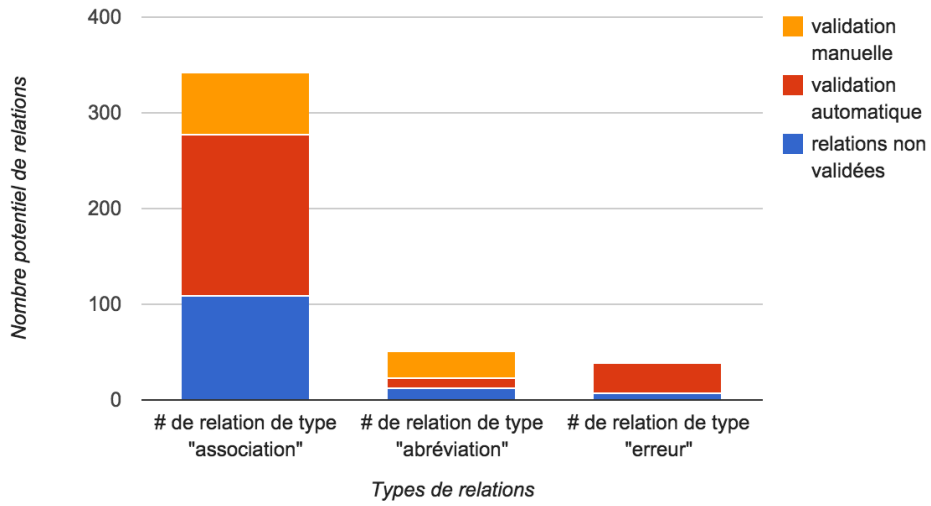

FIGURE 4 – Histogramme du nombre de relations validées automatiquement, manuellement et celles non validées.

6 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode permettant de relier les termes utilisés par les patients et constituant un CHV à ceux utilisés par les professionnels de santé et présents dans les vocabulaires contrôlés. Un avantage de cette méthode est qu'elle permet d'aligner des expressions pouvant être composées de plusieurs mots et de solliciter l'expert uniquement pour les termes pour lesquels il reste un doute (n'ayant pas été éliminés par le filtre automatique). Contrairement à la plupart des CHV existant uniquement en langue anglaise et construits manuellement, nous proposons une méthode semi-automatique originale pour construire un tel CHV pour le français. Nous avons appliqué cette méthode au domaine de la cancérologie mais elle peut être appliquée à de nombreux autres domaines. Une telle ressource sera une brique essentielle à l'exploitation automatiquement du contenu des médias sociaux dans le domaine médical.

La ressource est actuellement téléchargeable librement pour la communauté `http://www.lirmm.fr/~tapinzali/Ressources/VocPatMed`. Nous sommes actuellement en train de transformer cette ressource dans un format lisible par l'être humain et par l'ordinateur. Pour ce faire, nous allons donc créer une ontologie au format SKOS pour l'intégrer sur la plateforme BioPortal (Noy *et al.*, 2009). SKOS fournit le vocabulaire nécessaire pour définir les attributs d'un concept et les relations entre les concepts qui nous permettra de garder des informations sur la méthode d'obtention du terme patient (orthographe, abbréviation et association), le type de validation et éventuellement le type de relation identifiée automatiquement dans Wikipédia ou dans la ressource jeux de mots (e.g. synonyme). Un mapping sera fait entre notre ontologie au format SKOS et celles déjà existantes sur BioPortal puisque nous apparierions des termes patients à des termes de l'INCa tous présents dans le MeSH. Grâce à l'*Annotator* de BioPortal,

nous pourrons alors annoter n'importe quels types des textes issus des médias sociaux échangeant dans le domaine biomédical.

Les expérimentations réalisées vont également être complétées. Nous allons notamment étudier la relation entre le poids et les validations expertes, afin de définir un seuil à partir duquel il n'est plus intéressant de proposer une relation à l'expert. Actuellement, nous avons mesuré à quel point nos associations étaient justes en calculant la précision mais nous devons également nous intéresser au rappel pour mesurer combien de relations nous ne sommes pas capables d'identifier en repartant des mots non appariés de l'INCa par exemple. Nous souhaitons également appliquer cette méthode sur d'autres jeux de données réels pour enrichir notre ressource. Entre chaque mise à jour de la ressource, nous conservons une *Liste noire* de tous les relations rejetées. Nous comparerons le vocabulaire acquis dans les forums et celui acquis dans les réseaux de type Facebook. Les forums seraient ils plus intéressants car plus techniques ?

Le choix de la ressource Wikipédia est également discutable car n'entrant pas dans la catégorie des PAT en se situant entre le discours des praticiens et des médecins, ni scientifique, ni grand public. D'autres ressources doivent être envisagées pour le filtrage réalisé lors de l'étape 5 par exemple en traduisant un CHV anglais.

À plus long terme, nous envisageons de ré-exploiter les données utilisées pour étudier la qualité de vie des patientes atteintes d'un cancer du sein, et ainsi améliorer nos processus comme celui présenté dans (Opitz *et al.*, 2014). Nous pourrons mesurer l'impact de la ressource. De même, que donnerait notre méthode sur des médias sociaux en anglais pour étendre les CHV existants ? Nous allons également étudier l'évolution du vocabulaire des patients au cours du temps en utilisant un modèle de type LDA (Lattent Dirichlet Allocation). Il s'agit d'un modèle bayésien hiérarchique fondé sur une catégorie de modèles « topic models » et qui cherchent à découvrir des structures thématiques cachées dans des vastes archives de documents.

7 Remerciements

Ces travaux ont été financés par l'ANR SFIR (Semantic Indexing of French Biomedical Data Resources) et par le projet « Comparison of longitudinal analysis models of the health-related quality of life in oncology » (financement IRESP).

Références

BUSCALDI D. & ROSSO P. (2006). Mining knowledge from wikipedia for the question answering task. In *Proceedings of the International Conference on Language Resources and Evaluation*, p. 727–730.

CHERNOV S., IOFCIU T., NEJDL W. & ZHOU X. (2006). Extracting semantics relationships between wikipedia categories. volume 206 : Citeseer.

DOING-HARRIS K. M. & ZENG-TREITLER Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. volume 13 : JMIR Publications Inc.

FISCELLA K., MELDRUM S., FRANKS P., SHIELDS C. G., DUBERSTEIN P., MCDANIEL S. H. & EPSTEIN R. M. (2004). Patient trust : is it related to patient-centered behavior of primary care physicians ? volume 42, p. 1049–1055 : LWW.

GABRILOVICH E. & MARKOVITCH S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, p. 1606–1611.

HANCOCK J. T., TOMA C. & ELLISON N. (2007). The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 449–452 : ACM.

LAFOURCADE M. & JOUBERT A. (2012). Increasing long tail in weighted lexical networks. In *Cognitive Aspects of the Lexicon (CogAlex-III), COLING*, p.1̃6.

LOSSIO-VENTURA J. A., JONQUET C., ROCHE M. & TEISSEIRE M. (2014a). Biotex : A system for biomedical terminology extraction, ranking, and validation. In *Proceedings of the 13th International Semantic Web Conference (ISWC'14). Trento, Italy*.

LOSSIO-VENTURA J. A., JONQUET C., ROCHE M. & TEISSEIRE M. (2014b). Integration of linguistic and web information to improve biomedical terminology extraction. In *Proceedings of the 18th International Database Engineering & Applications Symposium*, p. 265–269 : ACM.

LOSSIO-VENTURA J. A., JONQUET C., ROCHE M. & TEISSEIRE M. (2014c). Yet another ranking function for automatic multiword term extraction. In *Advances in Natural Language Processing*, p. 52–64 : Springer.

MACLEAN D. L. & HEER J. (2013). Identifying medical terms in patient-authored text : a crowdsourcing-based approach. p. amiajnl–2012 : BMJ Publishing Group Ltd.

MEROLLI M., GRAY K. & MARTIN-SANCHEZ F. (2013). Health outcomes and related effects of using social media in chronic disease management : A literature review and analysis of affordances. volume 46, p. 957–969 : Elsevier.

NOY N. F., SHAH N. H., WHETZEL P. L., DAI B., DORF M., GRIFFITH N., JONQUET C., RUBIN D. L., STOREY M.-A., CHUTE C. G. *et al.* (2009). Bioportal : ontologies and integrated data resources at the click of a mouse. p. gkp440 : Oxford Univ Press.

OPITZ T., AZÉ J., BRINGAY S., JOUTARD C., LAVERGNE C. & MOLLEVI C. (2014). Breast cancer and quality of life : medical information extraction from health forums. In *Medical Informatics Europe*, p. 1070–1074.

PATERNOSTRE M., FRANCQ P., LAMORAL J., WARTEL D. & SAERENS M. (2002). Carry, un algorithme de désuffixation pour le français.

PONZETTO S. P. & STRUBE M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, p. 192–199 : Association for Computational Linguistics.

RAMESH B. P., HOUSTON T. K., BRANDT C., FANG H. & YU H. (2013). Improving patients' electronic health record comprehension with noteaid. In *MedInfo*, p. 714–718.

SADILEK A., KAUTZ H. A. & SILENZIO V. (2012). Modeling spread of disease from social interactions. In *ICWSM*.

WANG P., HU J., ZENG H.-J. & CHEN Z. (2009). Using wikipedia knowledge to improve text classification. volume 19, p. 265–281 : Springer.

WITTEN I. & MILNE D. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence : an Evolving Synergy, AAAI Press, Chicago, USA*, p. 25–30.

WU D. T., HANAUER D. A., MEI Q., CLARK P. M., AN L. C., LEI J., PROULX J., ZENG-TREITLER Q. & ZHENG K. (2013). Applying multiple methods to assess the readability of a large corpus of medical documents. In *MedInfo*, p. 647–651.

ZENG Q. T. & TSE T. (2006). Exploring and developing consumer health vocabularies. volume 13, p. 24–29 : BMJ Publishing Group Ltd.

ZENG Q. T., TSE T., DIVITA G., KESELMAN A., CROWELL J., BROWNE A. C., GORYACHEV S. & NGO L. (2007). Term identification methods for consumer health vocabulary development. volume 9 : Internet Healthcare Coalition.