

Formalisation semi-automatique d'un vocabulaire patient/médecin dédié au cancer du sein

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

▶ **To cite this version:**

Mike Donald Tapi Nzali, Jérôme Azé, Sandra Bringay, Christian Lavergne, Caroline Mollevi, et al.. Formalisation semi-automatique d'un vocabulaire patient/médecin dédié au cancer du sein. Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle, 2016, 30 (5), pp.533-556. 10.3166/ria.30.533-556 . hal-01583190

HAL Id: hal-01583190
https://hal.science/hal-01583190

Submitted on 3 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Formalisation semi-automatique d'un vocabulaire patient/médecin dédié au cancer du sein

Mike Donald Tapi Nzali [1,3,4], **Jérôme Azé** [1,4], **Sandra Bringay** [2,4], **Christian Lavergne** [2,3], **Caroline Mollevi** [5], **Thomas Opitz** [6]

1. *Université de Montpellier, France, mike-donald.tapi-nzali@univ-montp2.fr*
2. *Université Paul Valéry, Montpellier 3, France, {sandra.bringay, christian.lavergne}@univ-montp3.fr*
3. *Institut Montpelliérain Alexander Grothendieck, France, christian.lavergne@univ-montp2.fr*
4. *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, France, {mike-donald.tapi-nzali, jerome.aze, sandra.bringay}@lirmm.fr*
5. *Institut du Cancer Montpellier, Montpellier, France, caroline.mollevi@icm.unicancer.fr*
6. *Biostatistique et Processus Spatiaux, INRA Avignon, France, thomas.optiz@paca.inra.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 ou Facebook, nous décrivons la construction d'une ressource lexicale qui aligne le vocabulaire des patients à celui des professionnels de santé. Nous utilisons plusieurs méthodes prenant en compte les aspects linguistiques et statistiques proposées dans la littérature pour construire cette ressource et nous la transformons en une ontologie SKOS (Simple Knowledge Organization System). 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.*

ABSTRACT. *Nowadays, social media is increasingly used by patients and health professionals. Most often, the patients are lay in the medical field, they use slang, abbreviations, and their own vocabulary during their exchanges. In order to automatically analyze texts from social networks, we need a specific vocabulary. Considering a corpus of documents from messages from social media like forums and Facebook, we describe the construction of a lexical resource that*

Revue d'intelligence artificielle – n° 5/2016, 533-555

aligns the vocabulary of patients to that of health professionals. In order to build this resource and transform it into a SKOS ontology, we use several methods taking into account the linguistic and statistical aspects proposed in the literature. On the one hand, this work will improve information retrieval in health forums and on the other hand it will facilitate the development of statistical studies based on information extracted from these forums.

MOTS-CLÉS : *extraction d'information, médias sociaux, mesure statistique, ontologie, vocabulaire patient.*

KEYWORDS: *information extraction, social media, statistic-based measure, ontology, patient vocabulary.*

DOI:10.3166/RIA.30.533-555 © 2016 Lavoisier

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 (*Patient-Authored Text* – PAT), provenant des messages issus des médias sociaux de type forums ou Facebook. Pour apparier les termes patients et médecins, nous comparons trois approches : la première est basée sur la structure de l'encyclopédie universelle collaborative Wikipédia ; la seconde est basée sur le moteur de recherche généraliste Google et sur la co-occurrence des termes patients et médecins sur les textes du web ; la troisième est également basée sur les co-occurrences des termes patients et médecins capturées au travers des productions textuelles des patients.

Notre méthode a été expérimentée avec succès sur un jeu de données réelles dans le domaine du cancer du sein. Un sous-ensemble de relations obtenues a été validé automatiquement en utilisant la ressource collaborative du site « *www.JeuxDeMots .org* ». Une validation manuelle a été également réalisée par cinq personnes, dont un expert du domaine médical.

Cet article est organisé comme suit. Dans la section 2, nous motivons notre travail et donnons un état de l'art. 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 présentons la formalisation du vocabulaire sous la forme d'une ontologie SKOS (*Simple Knowledge Organization*

System). Dans la section 6, nous discutons les résultats issus des trois méthodes. Finalement, dans la section 7, 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 assurant 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 patients atteints d'un cancer du sein à partir des médias sociaux. Nous avons cherché à capturer et quantifier ce que les patients 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] ou 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 proposons donc dans ce travail 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

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/

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, 2011) ont proposé une méthode pour générer automatiquement des termes candidats à traiter par des humains pour inclusion dans un CHV. Cependant, ils n'apparient pas automatiquement les termes des patients à ceux des médecins comme nous allons le proposer dans ce travail. 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] qui est inclus dans l'UMLS. À notre connaissance, il n'y a pas de CHV en français. Toutefois, on trouve certains travaux proches, comme ceux de (Bouamor *et al.*, 2016). Ils ont proposé une approche basée sur l'apprentissage par transfert en entraînant un système avec des fonctionnalités non-lexicales sur des données en anglais, puis en l'appliquant au français.

Dans les médias sociaux (forums, Facebook), le volume des textes rédigés par les patients est de plus en plus important (MacLean, Heer, 2013). Si de tels PAT ne sont pas suffisamment précis pour des objectifs scientifiques, ils donnent en temps réel accès à de très nombreuses descriptions de l'expérience des patients, sur un large éventail de sujets. 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]. Des méthodes basées sur les PAT permettent de construire des vocabulaires patients. Dans leurs travaux, (MacLean, Heer, 2013) ont proposé une méthodologie de « crowdsourcing » pour relier des termes médicaux à des PAT. Dans des travaux encore plus récents, (Elhadad *et al.*, 2014) ont proposé une méthode permettant de générer un lexique pour la langue anglaise représentatif des termes utilisés dans des PAT par les membres d'un forum. De même, 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 le web pour faire des appariements. Tout d'abord, nous utilisons 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

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/Wikipedia:Accueil_principal

multilingue qui couvre de très nombreux domaines. La version française, en date du 15 février 2016 contient plus d'un million et demi d'articles. Les articles étant finement structurés, Wikipédia 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 Wikipédia. (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 Wikipédia pour déterminer la proximité sémantique entre les mots. D'autres travaux plus récents de (Hamon, Grabar, 2015) utilisent l'encyclopédie Wikipédia et des corpus multilingues anglais et français pour associer les terminologies anglaises et françaises aux ressources terminologiques ukrainiennes. 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. Nous utilisons également des mesures de co-occurrence plus classiques pour calculer un degré d'association entre des termes patients et médecins. Les mesures d'association de mots sont utilisées dans plusieurs domaines comme l'écologie (Dice, 1945), la médecine (Lu *et al.*, 2015) et le traitement du langage (Islam *et al.*, 2012). De telles mesures ont été récemment étudiées dans (Zadeh, Goel, 2013 ; Zheng *et al.*, 2015 ; Nalawade *et al.*, 2016 ; Lossio-Ventura *et al.*, 2016), telles que Dice, Jaccard, Overlap ou Cosine. Dans ce travail, nous utiliserons une mesure adaptée de la mesure de Jaccard qui compare le nombre d'apparitions de termes à apparier indépendamment puis ensemble dans les PAT produits par les usagers des médias sociaux. Une autre mesure pour calculer l'association entre les mots utilise le nombre de pages retournées par les moteurs de recherche Web. Cette mesure est appelée *Normalized Google Distance* (Cilibrasi, Vitanyi, 2007). Elle s'appuie sur le nombre de fois où les mots apparaissent indépendamment et ensemble dans les documents indexés par un moteur de recherche. Avec les trois mesures utilisées, nous voulons tirer profit de la ressource lexicale Wikipédia en utilisant la mesure Wikipédia, des données textuelles des patients en utilisant la mesure de Jaccard et des données fournies par le web en utilisant la mesure de Google.

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 choisi comme ressource de référence le dictionnaire 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 « ressource INCa ».

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



Figure 1. Extraction des termes patients (équivalent des termes médicaux) à partir des média sociaux

Étape 1 : développement du corpus de messages. Nous utilisons des messages issus des forums échangeant sur le cancer du sein et des groupes de paroles Facebook. Ces derniers 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 entre 2010 et 2014 par 1 389 membres 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 plus classiques provenant du forum « *cancerdusein.org* ». Nous avons récolté 17 000 messages provenant de 665 utilisateurs publiés entre 2010 et 2014. 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 été effectué sur ces messages (pas de correction automatique, ni de lemmatisation).

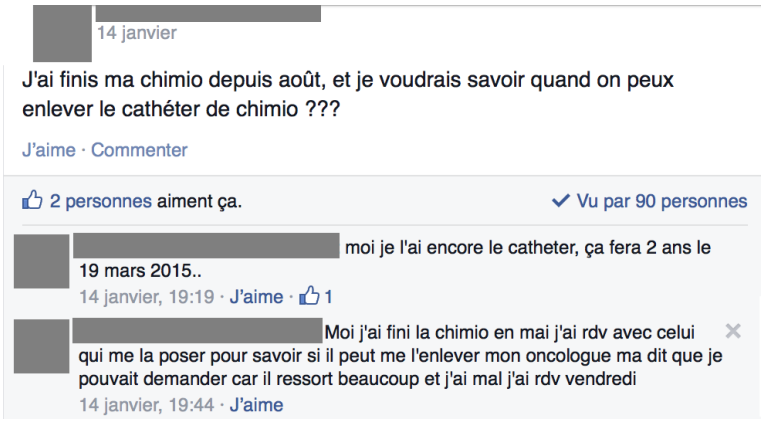


Figure 2. Posts anonymisés et commentés par des utilisateurs d'un groupe Facebook

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 statistiques 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 tableau 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 la ressource INCa, que nous allons utiliser dans les étapes 3, 4 et 5 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.

Tableau 1. Exemples de motifs linguistiques utilisés dans BioTex

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

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 cher-

chons à apparier tous les termes $t_i \in T$, avec un mot bien orthographié présent dans la ressource INCa. Pour cela nous utilisons le logiciel Aspell [11] pour obtenir un ensemble $M_i = \{m_1, m_2, ..., m_y\}$ de y propositions de corrections du mot t_i et ne conservons que les propositions présentes dans la ressource INCa. Nous utilisons ensuite la mesure de Levenshtein pour calculer la distance entre le terme t_i et chaque terme m_j ($j \in [1..Y]$). 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 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 $poids(m_j, t_i) = 1/|M_i|$. Le tableau 2 présente quelques fautes d'orthographe fréquemment rencontrées.

Tableau 2. Équivalent entre termes biomédicaux et termes patients (contenant des erreurs orthographiques)

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

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 $A_i = \{a_1, a_2, ..., a_k\}$ de k propositions d'abréviations incluses dans la ressource INCa. Le terme t_i est associé à une abréviation a_j avec un $poids(a_j, t_i) = 1/|A_i|$. Des exemples de termes appariés avec cette méthode sont listés dans la table 3.

Tableau 3. Équivalent entre termes biomédicaux et termes patients (abréviations)

Termes biomédicaux	**Termes patients**
oncologue	*onco*
chimiothérapie	*chimio*
mammographie	*mammo*

Étape 5 : similarité entre 2 mots. Nous nous intéressons ici à tous les termes produits à l'étape 2 qui ne sont ni des fautes d'orthographe fréquentes (repérées à l'étape 3), ni des abréviations (repérées à l'étape 4). Nous cherchons à apparier ces termes à

11. http://aspell.net/

trois niveaux : en considérant une ressource structurée sémantiquement (Wikipedia), en considérant des cooccurrences généralistes dans les textes du web avec le moteur de recherche Google et en considérant les cooccurrences dans les messages des patients avec la mesure de Jaccard.

– Mesure de similarité calculée à partir des pages Wikipédia. L'hypothèse ici est d'utiliser la structure sémantique des liens entre les pages de la ressource Wikipédia. Pour cela, nous interrogeons cette ressource grâce à son API [12]. Dans cette encyclopédie, un terme (*mot*) référencé est décrit par une page [13] 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 [14]. Certaines relations entre termes Wikipédia sont typées (e.g. synonymie). Sur la partie gauche de la figure 3, on retrouve la page du terme *Tumeur* et sur la partie droite, les termes liés. Soit $W_t = (w_1, ..., w_n), n \in \mathbb{N}^*$ l'ensemble des termes liés par Wikipédia à un terme t et appartenant à la ressource INCa. Un terme t est associé à un terme w_i selon une mesure calculée en utilisant la formule 1. Notons que l'ensemble ne contient que les termes présents dans la ressource INCa, ce qui nous assure de ne pas avoir des associations comme « Tumeur » et « Jules César » (voir figure 3).

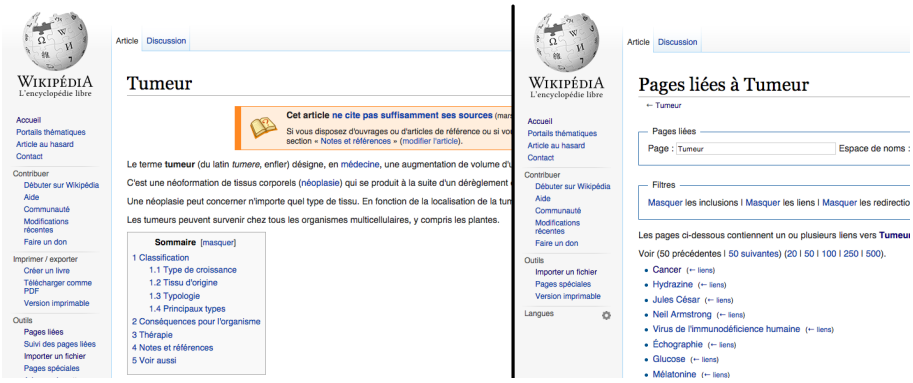


Figure 3. Page Wikipédia et page liée

$$Wiki(w_1, w_2) = \frac{MoyNW(w_1, w_2)}{\sum_{k=1}^{|W|} MoyNW(w_k, w_2)} \quad (1)$$

$$MoyNW(w_1, w_2) = \frac{NW(w_1, w_2) + NW(w_2, w_1)}{2} \quad (2)$$

où $NW(w_i, w_j)$ est la fréquence d'apparition du terme w_i dans la page Wikipédia du terme w_j.

12. http://fr.wikipedia.org/w/api.php?
13. *http://fr.wikipedia.org/wiki/**mot***
14. *http://fr.wikipedia.org/wiki/Spécial:Pages_liées/**mot***

– Mesure de similarité Google. L'hypothèse ici est d'exploiter les co-occurences dans les textes indexés par le moteur généraliste Google. Nous utilisons la mesure de similarité proposée par (Cilibrasi, Vitanyi, 2007). Il s'agit d'une mesure de similarité sémantique basée sur le nombre de résultats retournés par une requête Google entre deux termes. Cette distance normalisée est obtenue comme suit :

$$NGD(w_1, w_2) = \frac{max\{log\ NG(w_1), log\ NG(w_2)\}\ -\ log\ NG(w_1, w_2)}{log\ M - min\{log\ NG(w_1), log\ NG(w_2)\}} \quad (3)$$

où $NG(w_i)$ est le nombre de « hits » (pages retournées) de Google pour le terme w_i et $NG(w_i, w_j)$ est le nombre de « hits » pour le couple de mots w_i et w_j et M est le nombre de pages web indexées par Google.

– Coefficient de Jaccard. L'hypothèse ici est d'exploiter les co-occurences non plus sur le web mais dans les textes produits par les patients. En effet, nous avons remarqué que fréquemment, dans le cas de maladies chroniques comme le cancer, le patient utilise de l'argot puis au contact de la communauté s'approprie le vocabulaire des professionnels de santé jusqu'à parler comme eux. Si l'on considère l'ensemble de ses messages, on trouve souvent des mots d'argots et des termes médicaux associés. Nous utilisons une formule similaire à celle de Jaccard. Ici, nous cherchons à calculer la similarité entre w_1 et w_2 en utilisant le corpus C.

$$JAC(w_1, w_2) = \frac{NJ(w_1, w_2)}{NJ(w_1) + NJ(w_2) - NJ(w_1, w_2)} \quad (4)$$

$NJ(w_i)$ représente le nombre total d'apparitions du mot w_i dans une phrase du corpus C. $NJ(w_i, w_j)$ représente le nombre total de co-occurences dans une phrase des mots w_i et w_j. Nous considérons une phrase comme l'ensemble des messages d'un patient.

4. Résultats

À l'issu du processus précédent et indépendamment de la mesure utilisée, nous avons obtenu k relations r_i avec $i \in [1, k]$. Chaque relation r_i relie un mot patient pat_j [15] avec un mot médecin bio_l [16]. Chaque relation est associée à une méthode d'obtention $meth \in$ *{Erreur orthographique, Abréviation, Association}*. Dans cette section, nous présentons les deux méthodes de validation utilisées (automatique et manuelle) et les différents résultats obtenus. La validation finale manuelle est importante pour présenter les faiblesses des associations obtenues avec les méthodes quantitatives.

Nous évaluons nos résultats en termes de précisions obtenues sur les *k* premiers termes issus de l'étape 2 de la figure 1. Nous discutons ces résultats dans la section 6.

15. Les pat_j sont les termes issus du corpus.
16. Les bio_l sont les termes du dictionnaire fourni par l'INCa.

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. Ce choix est justifié par le critère strict que nous mettons sur la distance de Levenshtein dans notre algorithme de correction orthographique.

– La paire pat_j - bio_l existe dans le dictionnaire de relations fourni par le jeu 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 occurrences 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.

Tableau 4. Exemples de termes validés automatiquement en utilisant JeuxDeMots

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*
tumeur	*cancer*	*association*
chute des cheveux	*alopécie*	*association*

4.2. Validation manuelle

Toutes les relations r_i n'ayant pas pu être validées automatiquement sont présentées à cinq personnes, dont un expert du domaine médical pour validation manuelle [17]. Nous leur proposons des relations sous la forme : « *terme - terme associé - type de la relation* » afin de valider l'association et l'étiquette. Deux choix sont proposés : 1) **Oui** : pour valider la relation ; 2) **Non** : pour invalider la relation. Nous conservons une relation si au moins trois annotateurs sur les cinq l'ont validé.

Des exemples de relations validées manuellement sont présentés dans le tableau 5.

17. Une image de l'interface de validation est présente à cette url: *http://www.lirmm.fr/~tapinzali/Validation/Validation.php*

Tableau 5. Exemples de termes validés manuellement

Terme patient	Terme biomédical	Relation
psy	*psychologue*	*abréviation*
onco	*oncologue*	*abréviation*
gynéco	*gynécologue*	*abréviation*
constipation	*laxatif*	*association*
libido	*sexologie*	*association*
morphine	*douleur*	*association*
huile de ricin	*laxatif*	*association*

4.3. Validation globale

Comme résultat, nous obtenons un ensemble de relations (pat_j - bio_l). Ne sachant pas à l'avance combien de relations il existe pour un mot patient pat_j, nous ne pouvons faire une évaluation en terme de rappel. Pour cela, nous avons décidé comme (Doing-Harris, Zeng-Treitler, 2011) d'évaluer nos résultats en termes de précision. Nous utilisons la formule 5.

$$P = \frac{|R_a| + |R_m|}{|R|} \; et \; R_a \cap R_m = \emptyset, R_a \subseteq R, R_m \subseteq R, |R| \geq |R_a| + |R_m| \quad (5)$$

où R_a est l'ensemble des relations validées automatiquement, R_m est l'ensemble des relations validées manuellement et R est l'ensemble des relations ayant été fournies en sortie par notre outil.

5. Formalisation de la ressource sous la forme d'une ontologie en SKOS

Afin de bénéficier des avantages des vocabulaires contrôlés, nous avons formalisé le vocabulaire précédent sous la forme d'une ontologie. Les applications typiques de vocabulaires contrôlés sont la classification, l'indexation, l'auto-complétion, la reformulation de requêtes, etc. Nous avons décidé d'utiliser SKOS (Simple Knowledge Organization System). SKOS est un langage de représentation de schémas de concepts tels que les thésaurus, les taxonomies et les vocabulaires contrôlés (Miles, Bechhofer, 2005). Il permet d'exprimer et de gérer très simplement des modèles interprétables par les machines dans la perspective du web sémantique. SKOS étant lui-même une ontologie OWL, sa représentation de SKOS repose sur des graphes RDF.

De plus en plus de vocabulaires sont implémentés en SKOS. (Van Assem *et al.*, 2006) ont proposé une méthode pour passer d'un thésaurus à une ontologie SKOS et l'ont appliquée aux domaines de la santé et de l'audiovisuel. (Solomou, Papatheodorou, 2010) ont converti un thésaurus de termes grecs en ontologie pour le domaine éducatif et culturel. (Summers *et al.*, 2008) présentent une technique pour convertir

des données MARCXML [18] en une ontologie SKOS et l'appliquent sur les données du web. Il existe des versions SKOS de thésaurus très utilisés comme Agrovoc [19] pour l'alimentation et l'agriculture, Eurovoc [20] pour décrire les activités de l'Union Européenne, GEMET [21] pour l'environnement et STW [22] pour l'économie.

Dans ce travail, nous utilisons le vocabulaire présenté à la section 4 pour créer une ontologie SKOS. Cette ontologie permettra de lier un terme de la ressource INCa à différents mots patients : des termes préférentiels pour définir le terme MeSH ; des termes alternatifs seront dans notre cas des synonymes, des abréviations, ... ; des termes cachés pour représenter les fautes d'orthographes. Un même terme patient pourra être associé à plusieurs termes médecin (e.g. *onco* pour *oncologue* ou *oncologie*).

Un extrait du graphe proposé pour représenter l'ontologie est présenté dans la figure 4 pour le concept *« chimiothérapie »*.

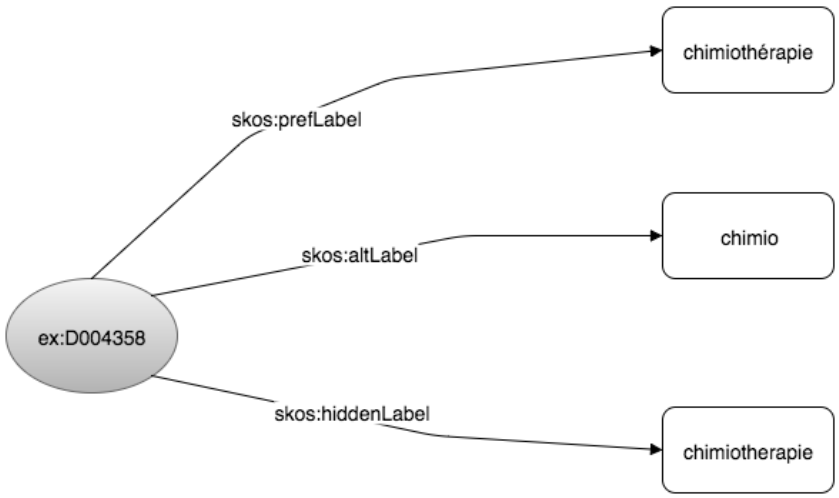


Figure 4. Extrait de l'ontologie SKOS pour le terme « chimiothérapie »

6. Discussion

Après validation automatique des résultats obtenus grâce aux différentes mesures utilisées, les validations manuelles de nos relations ont été calculées par consensus entre les différents annotateurs. Cinq annotateurs ont participé aux annotations, dont un expert du domaine médical. Un coefficient de kappa de Fleiss kf a été calculé pour

18. http://www.loc.gov/standards/marcxml/
19. http://aims.fao.org/website/AGROVOC-Thesaurus/sub
20. http://eurovoc.europa.eu/
21. http://www.eionet.europa.eu/gemet
22. http://zbw.eu/stw/versions/latest/about

mesurer l'accord inter-annotateur. Nous obtenons un kf égal à 0,25 (accord faible, dû à la variabilité individuelle du jugement des annotateurs sur l'intérêt médical des termes) pour les mesures Wikipédia et Google. Cet accord pourrait être amélioré par discussion des guides d'annotation ou par une phase de réconciliation parmi les annotateurs.

6.1. Comparaison des mesures

Nous avons utilisé trois mesures (Wikipédia, Google et Jaccard) pour calculer la similarité entre les termes des relations candidates identifiées à l'étape 4 et avons appliqué la méthodologie décrite dans la section 4 pour les comparer. Nous discutons dans la suite des résultats obtenus avec 1 900 candidats évalués, il s'agit des 1 900 premiers termes renvoyés par BioTex lors de l'étape 2. Les figures 5 et 6 présentent les résultats obtenus selon la méthode de validation des relations sur les deux corpus.

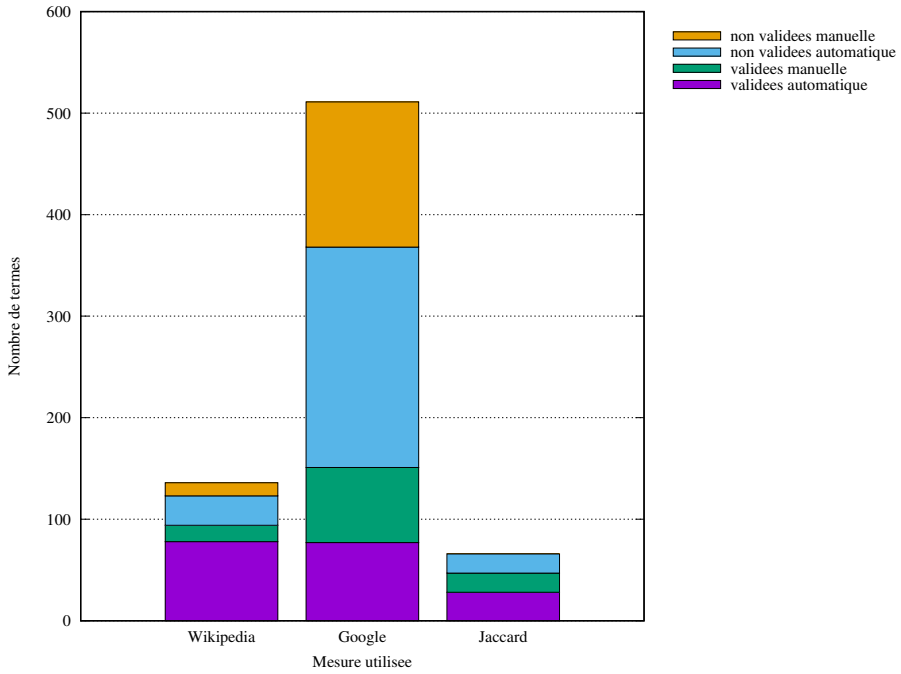


Figure 5. Histogramme du nombre de relations validées automatiquement, manuellement et celles non validées pour toutes les mesures sur le corpus « cancerdusein.org »

– **Mesure de Wikipédia** : sur les 1 900 premiers termes que nous avons traités, nous avons obtenu une précision globale *P* de 88 % (respectivement 91 %). Sur le

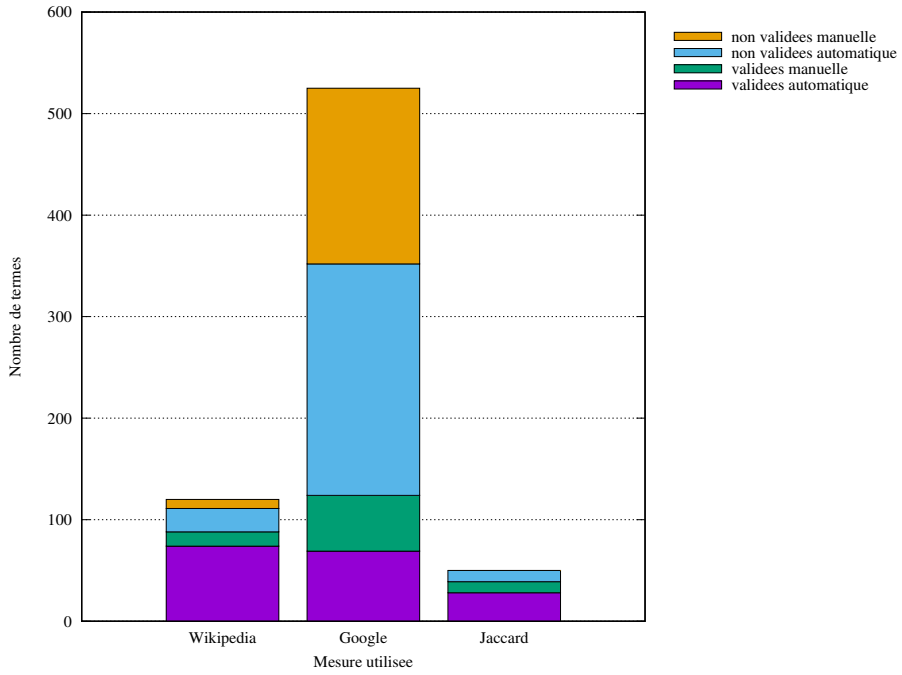


Figure 6. Histogramme du nombre de relations validées automatiquement, manuellement et celles non validées pour toutes les mesures sur le corpus « des groupes Facebook »

corpus « cancerdusein.org », nous avons validé 94 relations sur les 107 relations obtenues, dont 78 relations obtenues par validation automatique et 16 par validation manuelle. Sur le corpus « Facebook », nous avons validé 88 relations sur les 97 relations obtenues, dont 74 relations obtenues par validation automatique et 14 par validation manuelle. La mesure Wikipédia présente l'avantage de ramener très peu de bruit et donc peu de travail pour l'expert.

– **Mesure de Google** : sur les 1 900 premiers termes que nous avons traités, nous avons obtenu respectivement sur le corpus « cancerdusein.org » et « Facebook » une précision globale P de 52 % (respectivement 43 %). Sur le corpus « cancerdusein.org », nous avons validé 151 relations sur les 288 relations obtenues dont 77 relations obtenues par validation automatique et 74 par validation manuelle. Sur le corpus « Facebook », nous avons validé 124 relations sur les 290 relations obtenues dont 69 relations obtenues par validation automatique et 55 par validation manuelle. La mesure de Google est celle qui donne le plus de relations validées. Par contre, elle nécessite un effort supplémentaire de la part de l'expert en terme de validation manuelle car elle ramène beaucoup plus de bruit.

– **Mesure de Jaccard** : sur les 1 900 premiers termes que nous avons traités, nous avons obtenu respectivement sur le corpus « cancerdusein.org » et « Facebook » une précision globale P de 100 % (respectivement 100 %). Sur le corpus « cancerdusein.org », nous avons validé 47 relations sur les 47 relations obtenues dont 28 relations obtenues par validation automatique et 19 par validation manuelle. Sur le corpus « Facebook », nous avons validé 39 relations sur les 39 relations obtenues dont 28 relations obtenues par validation automatique et 11 par validation manuelle. Cette mesure est très intéressante car même si elle ramène peu de relations, toutes sont validées automatiquement, avec peu de travail de l'expert. Ceci se justifie par le fait qu'on a une base de recherche très restreinte avec cette mesure, car on utilise juste le corpus de documents des patients.

Nous nous sommes également posé la question du nombre de relations candidates à considérer à l'issue de l'étape 5. Selon la mesure utilisée, le constat diffère comme le montre les figures 7 et 8. Ces figures présentent les précisions (validation automatique et manuelle) des k premiers termes pour les trois mesures présentées dans la section 3. Nous remarquons qu'en utilisant la mesure de Google, plus le nombre de termes à traiter augmente, plus la précision diminue. La mesure Wikipédia quant à elle est plus stable. Quant à la mesure de Jaccard, bien que le nombre de relations trouvées soit petit, ces dernières sont toutes validées.

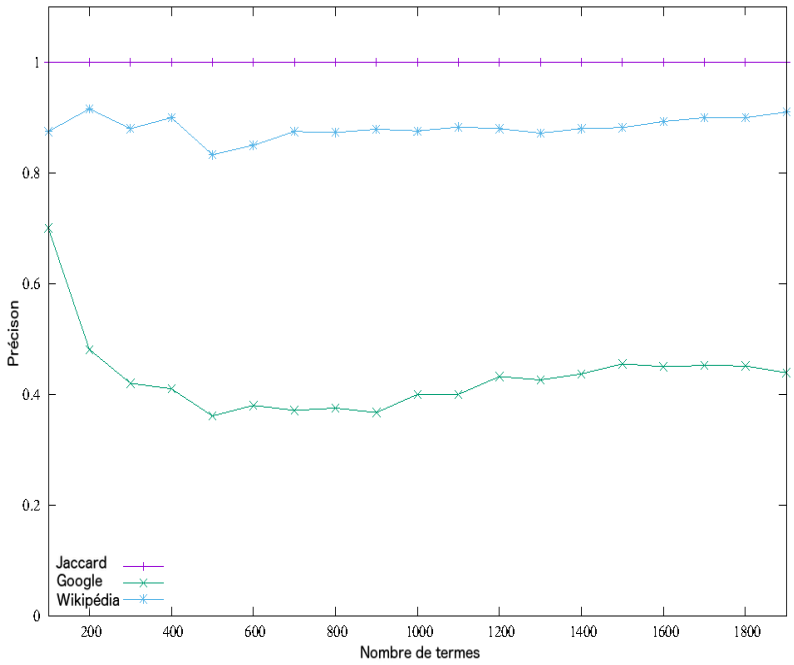


Figure 7. Précisions selon le nombre de termes choisis obtenues par combinaison des validations automatiques et manuelles sur le corpus du forum « cancerdusein.org »

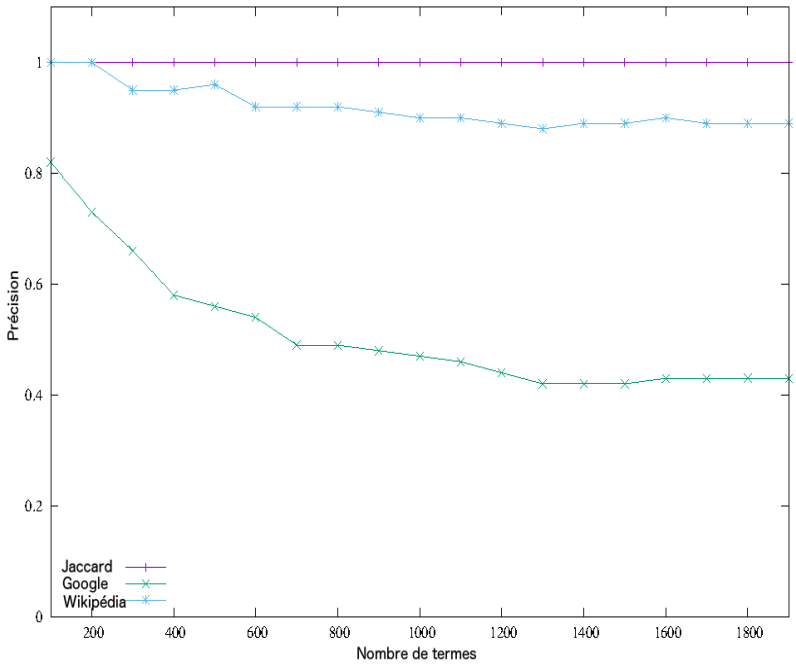


Figure 8. Précisions selon le nombre de termes choisis obtenues par combinaison des validations automatiques et manuelles sur le corpus « des groupes Facebook »

Au final, sur les k (k = 1 900) premiers termes traités dans chaque corpus (voir figure 9), nous obtenons au total 218 relations, dont 22 relations de type « erreur orthographique », 5 relations de type « abréviation » et 191 relations de type « association ». Cette ressource est actuellement téléchargeable librement pour la communauté à l'adresse suivante : *http://www.lirmm.fr/~tapinzali/Ressources/VocPatMed* sous forme plate et à l'adresse *http://bioportal.lirmm.fr/ontologies/MUEVO* sous forme d'ontologie SKOS. Par ailleurs, le nombre limité de relations totales obtenues dans notre cas s'explique également par le nombre de termes médecin cibles auxquels nous cherchons à apparier les termes des patients (ceux de la ressource INCa qui contient uniquement 1 227 termes). En effet, nous cherchons à créer une ressource spécifique au cancer du sein et non une ressource généraliste. Étant donné les relations obtenues par les mesures Wikipédia et Google sur les corpus utilisés, nous pouvons remarquer une certaine complémentarité des deux mesures, car elles retrouvent des relations différentes. Finalement, sur ces données textuelles bruitées biomédicales extraites des forums de santé, les résultats que nous obtenons sont très encourageants. (Doing-Harris, Zeng-Treitler, 2011) a réalisé un travail qui est très proche du notre, mais avec un objectif plus généraliste qui est celui de créer un CHV en langue anglaise. Sur 88 994 termes, ils ne trouvent que 774 relations et n'en valident que 237, soit une précision de 31 %.

6.2. Comparaison des corpus

Un constat est fait sur les résultats des deux mesures (Wikipédia et Google) : sur les données des groupes Facebook, on trouve 49 relations communes, 75 spécifiques à la mesure Google et 39 spécifiques à la mesure Wikipédia (voir figure 11). Sur les données des forums, on trouve 53 relations communes, 98 spécifiques à la mesure Google et 41 spécifiques à la mesure Wikipédia (voir figure 10).

Nous avons obtenu au total 192 relations sur le corpus du forum « cancerdusein.org », et 163 relations sur le corpus « Facebook ». On trouve 145 relations communes dans les deux corpus, soit 75 % de relations du corpus « cancerdusein.org » dans le corpus « Facebook » et 89 % des relations du corpus « Facebook » dans le corpus du forum « cancerdusein.org » (voir figure 9). Par ailleurs, nous retrouvons exactement les mêmes relations de type « abréviation » et « erreur orthographique ». Les relations qui diffèrent dans les deux corpus sont celles de type « association ». Vu le pourcentage des relations communes obtenues dans les deux corpus, une première remarque serait de dire que les patients utilisent un vocabulaire semblable dans les deux types de médias sociaux sur lesquels nous avons effectué nos expérimentations.

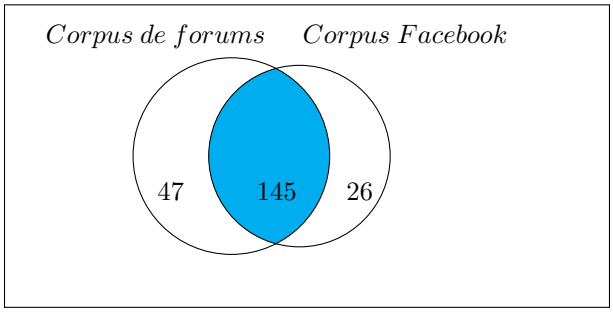


Figure 9. Diagramme de Venn des relations validées sur le corpus du forum et le corpus des groupes Facebook

Sur les données des forums, en validation automatique, la mesure Wikipédia donne de meilleurs résultats que celle de Google avec plus de relations soit 78 relations validées sur 107 candidates (72 %) alors que la mesure de Google valide 77 relations sur 288 relations candidates (27 %). Sur les données provenant de Facebook, les observations sont les mêmes, la mesure Wikipédia donne de meilleurs résultats. On obtient 74 relations validées sur 97 candidates (76 %) avec la mesure Wikipédia et 69 relations validées sur 290 candidates (24 %). Une deuxième remarque est que les deux mesures ont le même comportement sur les deux types de corpus.

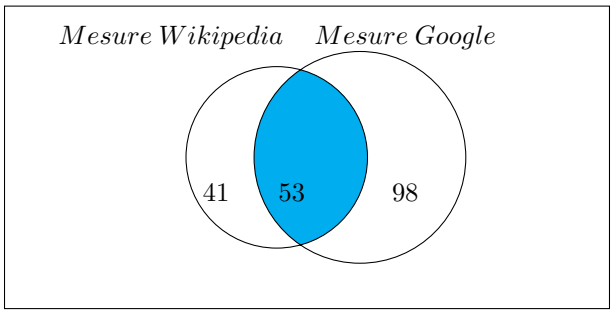


Figure 10. Diagramme de Venn des 192 relations validées sur les données du forum

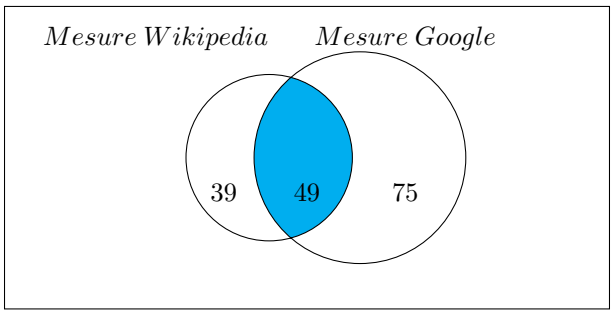


Figure 11. Diagramme de Venn des 163 relations validées sur les données Facebook

7. 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. Nous avons construit un vocabulaire patient/médecin à partir de forums de santé et de groupes Facebook et l'avons mis sous forme d'ontologie SKOS. Cette ressource pourra être mise à jour à chaque découverte d'une nouvelle relation (patient/médecin). Dans la ressource initiale de l'INCa, composé de 1 227 termes, 117 termes biomédicaux ont trouvé des correspondants avec des termes patients avec notre méthode (qui correspondent à 10 % de de la ressource initiale). Ceci peut être expliqué par le fait que nous ne considérons pas les 470 acronymes (qui correspondent à 38 % de la ressource initiale). De plus, nous avons projeté les 640 termes qui n'ont pas trouvé de correspondants patients dans le corpus et nous avons remarqué que certains parmi eux sont fréquemment utilisés par les patients et n'ont donc pas de substituts spécifiques aux patients. Dans la ressource produite, un terme peut-être lié à plusieurs termes patient (e.g. *onco — oncologue* et *onco – oncologie).*

Un avantage de la méthode proposée dans ce travail 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 automatique du contenu des médias sociaux dans le domaine médical.

Nous avons comparé trois mesures permettant d'apparier des termes patients et médecins. La mesure Wikipédia nécessite moins de travail pour l'expert que la mesure de Google pour laquelle la majorité des relations se valident manuellement. La mesure Jaccard est moins intéressante car elle renvoie très peu de relations et toutes sont trouvées par les mesures Wikipédia et Google. En termes de nombre de relations validées au final, les résultats obtenus par la mesure Google sont légèrement supérieurs à ceux obtenus par la mesure Wikipédia. Toutefois, au vu des précisions obtenues, la mesure préconisée pour limiter le travail de l'expert du domaine est la mesure de Wikipédia.

La ressource est actuellement téléchargeable librement pour la communauté à l'adresse suivante : *http://bioportal.lirmm.fr/ontologies/MUEVO*. Nous avons transformé cette ressource dans un format lisible par l'être humain et par l'ordinateur. Pour ce faire, nous avons créé 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. Ceci nous permettra de garder des informations sur la méthode d'obtention du terme patient (orthographe, abréviation et association), du type de validation et éventuellement du type de relation identifiée automatiquement dans Wikipédia ou dans la ressource jeux de mots (e.g. synonyme). Dans sa version actuelle, l'ontologie est très simple dans ce sens où elle est créée juste avec la relation obtenue dans le fichier retourné après validation automatique et manuelle. Nous envisageons donc de faire un étiquetage plus fin de nos relations et ainsi d'enrichir l'ontologie avec des informations telles que la provenance de la relation (quelle mesure, quel corpus, quel mode de validation, etc.) et également fournir un indice de confiance dans la relation.

À 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, par exemple sur les tâches d'annotation et de classification. De même, nous nous interrogerons sur l'intérêt de 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). Nous utiliserons donc le vocabulaire construit dans la phase de prétraitements de nos textes en remplaçant tous les termes patients par leur correspondant biomédicaux et prendrons le nouveau corpus en sortie pour appliquer au modèle LDA. Il s'agit d'un modèle bayésien hiérarchique fondé sur une catégorie de modèles « *topic model* » et qui cherchent à découvrir des structures thématiques cachées dans des vastes archives de documents.

Remerciements

Ces travaux ont été financés par l'ANR SFIR (Semantic Indexing of French Biomedical Data Resources) et par par l'Institut de Recherche en Santé Publique (http://www.iresp.net).

Bibliographie

Bouamor D., Llanos L. C., Ligozat A.-L., Rosset S., Zweigenbaum P. (2016). Transfer-based learning-to-rank assessment of medical term technicality. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, p. 2312–2316.

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. *Semantic Wiki*, vol. 206, p. 153-163.

Cilibrasi R. L., Vitanyi P. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, n° 3, p. 370–383.

Dice L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, vol. 26, n° 3, p. 297–302.

Doing-Harris K. M., Zeng-Treitler Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of Medical Internet Research*, vol. 13, n° 2, p. e37.

Elhadad N., Zhang S., Driscoll P., Brody S. (2014). Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *American Medical Informatics Association, Annual Symposium*, p. 516-525.

Fiscella K., Meldrum S., Franks P., Shields C. G., Duberstein P., McDaniel S. H. *et al.* (2004). Patient trust: is it related to patient-centered behavior of primary care physicians? *Medical Care*, vol. 42, n° 11, p. 1049–1055.

Gabrilovich E., Markovitch S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. *International Joint Conference on Artificial Intelligence*, vol. 7, p. 1606–1611.

Hamon T., Grabar N. (2015). Acquisition of medical terminology for ukrainian from parallel corpora and wikipedia. In *Terminologie Intelligence Artificielle*, p. 71-79.

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.

Islam A., Milios E. E., Keselj V. (2012). Comparing Word Relatedness Measures Based on Google n-grams. In *International Conference on Computational Linguistics*, p. 495-506.

Lafourcade M., Joubert A. (2012). Increasing long tail in weighted lexical networks. In *Cognitive Aspects of the Lexicon, International Conference on Computational Linguistics*, p. 5-20.

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 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, p. 157–160.

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.

Lossio-Ventura J. A., Jonquet C., Roche M., Teisseire M. (2014c). Yet another ranking function for automatic multiword term extraction. In *International Conference on Natural Language Processing*, p. 52–64.

Lossio-Ventura J. A., Jonquet C., Roche M., Teisseire M. (2016). Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, vol. 19, n° 1-2, p. 59–99.

Lu K., Mao J., Li G. (2015). Enhancing subject metadata with automated weighting in the medical domain: A comparison of different measures. In *International Conference on Asian Digital Libraries*, p. 158–168.

MacLean D. L., Heer J. (2013). Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the American Medical Informatics Association*, vol. 20, n° 6, p. 1120–1127.

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. *Journal of Biomedical Informatics*, vol. 46, n° 6, p. 957–969.

Miles A., Bechhofer S. (2005). Skos simple knowledge organization system reference. In *W3C Recommendation, World Wide Web Consortium, http://www.w3.org/TR/skos-reference/, consulté le 18 février 2016.* Consulté sur http://www.w3.org/TR/skos-reference/ ,18August2009

Nalawade R., Samal A., Avhad K. (2016). Improved similarity measure for text classification and clustering. In *International Research Journal of Engineering and Technology*, p. 214–219.

Noy N. F., Shah N. H., Whetzel P. L., Dai B., Dorf M., Griffith N. *et al*. (2009). Bioportal: ontologies and integrated data resources at the click of a mouse. In *Nucleic Acids Research*, p. 170-173. 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. *Technical report, Paul Otlet Institute, 15 pages*.

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.

Ramesh B. P., Houston T. K., Brandt C., Fang H., Yu H. (2013). Improving patients' electronic health record comprehension with noteaid. In *World Congress on Health and Biomedical Informatics*, p. 714–718.

Sadilek A., Kautz H. A., Silenzio V. (2012). Modeling spread of disease from social interactions. In *International Conference on Weblogs and Social Media*, p. 322–329.

Solomou G., Papatheodorou T. (2010). The use of SKOS vocabularies in digital repositories: the DSpace case. In *International Conference on Semantic Computing*, p. 542–547.

Summers E., Isaac A., Redding C., Krech D. (2008). Lcsh, skos and linked data. In *International Conference on Dublin Core and Metadata Applications*, p. 25-33.

Van Assem M., Malaisé V., Miles A., Schreiber G. (2006). A method to convert thesauri to skos. In *European Semantic Web Conference*, p. 95-109. Springer.

Wang P., Hu J., Zeng H.-J., Chen Z. (2009). Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, vol. 19, n° 3, p. 265–281.

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. *et al.* (2013). Applying multiple methods to assess the readability of a large corpus of medical documents. In *World Congress on Health and Biomedical Informatics*, p. 647–651.

Zadeh R. B., Goel A. (2013). Dimension independent similarity computation. *The Journal of Machine Learning Research*, vol. 14, n° 1, p. 1605–1626.

Zeng Q. T., Tse T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, vol. 13, n° 1, p. 24–29.

Zeng Q. T., Tse T., Divita G., Keselman A., Crowell J., Browne A. C. *et al.* (2007). Term identification methods for consumer health vocabulary development. *Journal of Medical Internet Research*, vol. 9, n° 1, p. e4.

Zheng Y., Mobasher B., Burke R. (2015). Integrating context similarity with sparse linear recommendation model. In *International Conference on User Modeling, Adaptation, and Personalization*, p. 370–376.