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

Les vocabulaires contrôlés (e.g. SNOMED, MeSH, UMLS) jouent un rôle clé dans les applications biomédicales de fouille de textes. Ils 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. 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 ce travail, 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. D'une part, elle a été validée automatiquement en utilisant la ressource collaborative du site « JeuxDeMots.org » et d'autre part, avec une validation manuelle réalisée par 4 personnes, dont un expert du domaine du cancer du sein.

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. 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.

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

1. HON (Health On the Net) How Do General Public Search Online Health Information ? Avril 2011

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 ce travail de construire un vocabulaire dédié aux « consommateurs de soins de santé » (Consumer Health Vocabularies - CHV).

L'un des apports majeur de notre approche est d'utiliser l'architecture de l'encyclopédie Wikipédia[5] pour rapprocher des termes utilisés par les patients et des termes utilisés par des professionnels de la santé.

Nous proposons une méthode 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 vocabulaire donné sur le site de l'INCa[6] composé de 1 227 termes, tous présents dans le MeSH en version française, que nous noterons INCa.

Les 5 étapes proposées sont les suivantes :

1. **Développement du corpus de messages.** Nous utilisons des messages issus du réseau social Facebook et de forums échangeant sur le cancer du sein.
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[7].
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 apparions tous les termes détectés mal orthographié et les mettons en relation avec leur correspondant bien orthographié présent dans l'INCa. ‘
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.
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[8].

Quelques exemples d'associations détectées sur les données du site « cancer du sein.org » sont : correction orthographique : cirose – cirrhose, abcé – abcès ; inclusion : chimio – chimiothérapie, onco – oncologue ; termes co-occurrents : crabe – cancer, bouton – acné. Pour valider nos résultats, nous procédons, d'une part, à une validation automatique partielle en utilisant une ressource lexicale appariant des termes utilisés par le grand public, donc non spécialisé en santé, et d'autre part, à une validation manuelle qui sera faite par des oncologues spécialistes du cancer du sein.

Dans ce travail nous proposons une approche permettant de relier les termes utilisés par les patients aux termes utilisés par les professionnels de santé. Cette ressource est une brique essentielle pour exploiter automatiquement le contenu des forums de santé.

---

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/
5. http://fr.wikipedia.org/wiki/Wikipédia:Accueil_principal
6. http://www.e-cancer.fr/cancerinfo/ressources-utiles/dictionnaire/
7. http://tubo.lirmm.fr/biotex/
8. http://fr.wikipedia.org/w/api.php?