[Young Researcher Category]

La technicité des termes : le *v-tech* comme paramètre d'évaluation

1. Introduction

Cette proposition s'inscrit dans le contexte de la *Terminologie Computationnelle* comme domaine d'étude récent qui vise à adopter des méthodes computationnelles et quantitatives afin de mener des recherches terminologiques et qualitatives [3, 5, 7]. L'état de l'art sur ce sujet montre que de nombreuses études (pour ne citer que les plus récentes : [1, 2, 15, 17]) sont principalement axées sur l'*extraction automatique de termes* à partir d'un corpus de documents et proposent différentes approches 1) linguistique, 2) statistique et 3) hybride.

L'acquisition de termes liés et pertinents à un domaine spécifique de l'activité humaine est effectuée automatiquement à l'aide d'approches quantitatives importées par le domaine de la recherche d'information (RI) : *Term frequency/Inverse Document Frequency* (TF/IDF), *Information Mutuelle*, *T-Score, C/NC value*, etc.[1] En outre, des ressources spécifiquement conçues pour cette tâche ont été réalisées afin d'augmenter les performances d'extraction : TermoStat [4], BiTermEx [13], TermEvaluator [10] et le récent projet MultiMedica [15] pour l'acquisition des termes concernant le domaine médical. L'importance d'une résolution efficace de cette tâche se reflète enfin dans de nombreux domaines de recherche. L'extraction automatique de termes permet d'effectuer des tâches liées à la recherche d'information (comme le repérage de documents pertinents pour une requête donnée), à la fouille de textes (*text mining*), à la création de ressources terminologiques, etc.

Le point de départ de tous ces travaux se reflète, tout d'abord, dans l'identification de termes qui font l'objet de l'extraction et par conséquent dans la distinction entre, d'une part, les mots d'ordre général et d'autre part, les termes spécifiques d'un domaine donné. En effet, toutes les études précédemment citées portent (plus ou moins explicitement) sur le concept de « poids » des termes dans une collection des documents afin d'indiquer les différents degrés de pertinence à un domaine. Ce concept a été exprimé, au fil du temps, à travers différentes dénominations. En 1972, Karen Spärck Jones [18] définissait la notion de « spécificité » des termes, une valeur calculable en fonction de la fréquence d'apparition des termes dans une collection des documents. En 1996, Kageura et Umino [11] introduisaient le concept de « termhood » (termicité ou termitude [9]) afin d'indiquer le degré de relation d'une unité linguistique à des concepts spécifiques pour un domaine, une valeur qui pouvait également être calculée à l'aide d'approches statistiques. Dans ce sens, le degré de termicité d'un terme est donc une valeur déterminante pour la tâche d'extraction automatique des termes et repose, en général, sur la fréquence d'apparition d'un terme candidat dans le corpus analysé.

Dans ce contexte, notre étude vise à repenser le concept de poids d'un terme en se concentrant plutôt sur la notion de « technicité », non dans sa connotation négative de difficulté de compréhension,[2] mais comme propriété sémantique des termes eux-mêmes. Nous nous proposons donc de donner une nouvelle acception à ce concept et de formaliser cette idée à travers une valeur attribuable selon le degré d'association avec le domaine d'intérêt. Pour cette raison, notre proposition ne vise pas à fournir une nouvelle méthode d'extraction automatique de termes, mais plutôt à définir un nouveau paramètre pour leur évaluation.

2. La valeur de technicité d'un terme : le *v-tech*

Dans les travaux susmentionnés, le poids d'un terme est déterminé par la fréquence et la distribution de ses occurrences dans un document et/ou un ensemble de documents. Pour cette raison, cette valeur est strictement dépendante et variable en fonction du corpus analysé. Notre étude vise plutôt à considérer le poids d'un terme, dans le sens du degré de « technicité », comme une propriété intrinsèque au terme lui-même et pas comme une valeur statistique liée au corpus de documents.

En définissant cette propriété par le critère de « technicité », nous ne faisons pas référence aux concepts de 1) « spécificité » des termes, comme valeur liée à la fréquence, 2) « termicité », comme propriété d'être ou ne pas être un terme, 3) « difficulté » de compréhension de la part d'un destinataire.

La valeur de « technicité » que nous proposons ici dépend du degré d'association du terme à un domaine d'intérêt. Pour donner une explication intuitive, nous prenons en considération les termes « quadrantectomie » et « patient » qui relèvent du domaine médical. Les deux termes semblent avoir un poids et un degré

[1] Pour une synthèse des principaux critères voire le tableau de [14].

[2] Voir par exemple les nombreuses études relatives à la difficulté de compréhension de la terminologie médicale dans le dialogue médecin-patient [2, 7, 12, 19].

de technicité différent : le terme « quadrantectomie »[3] désigne un concept unique dans le domaine de la chirurgie et il est répandu entre les spécialistes du domaine, alors que le terme « patient »[4] véhicule des significations différentes dans plusieurs domaines (médecine, philosophie et linguistique) et son usage est plus répandu, même dans le langage courant. En excluant de cette analyse la diffusion, en tant que concept proche de la fréquence d'apparition d'un terme dans un corpus réel ou imaginaire, nous nous concentrons sur la définition de « technicité » comme propriété dont la valeur est inversement proportionnelle au nombre de domaines dans lesquels un terme apparaît. En ce sens, plus un terme est polysémique, c'est-à-dire appartenant à plusieurs domaines, moins il sera technique car il ne sera ni mono-référentiel ni exclusif. En fait, une caractéristique souhaitable de la terminologie employée pour un domaine est celle d'être mono-référentielle et spécifique [16, 8] : la relation entre signe et référent devrait être univoque afin d'éviter les ambiguïtés et la polyvalence du point de vue sémantique. Or, un signe linguistique/terme qui désigne plus de référents perd de « technicité », car il n'est plus exclusif pour un seul domaine.

À partir de cette définition, nous avons formalisé ce principe en attribuant une valeur numérique à la technicité d'un terme. Nous introduisons donc la valeur que nous appellerons *v-tech* par la formule suivante :

$$v\text{-}tech\ (t) = k / d_t$$

La valeur *v-tech* pour un terme *t* est inversement proportionnelle à d_t, c'est-à-dire le nombre de domaines dans lesquels un terme apparaît, et *k* est un paramètre qui adapte la rapidité avec laquelle le terme *t* perd en technicité avec l'augmentation du nombre des domaines. Le principe à la base de *v-tech* est que les termes polysémiques, ayant plusieurs domaines d'utilisation, auront une valeur qui tendra à 0. L'image suivante illustre l'évolution de *v-tech* en fonction de d_t lorsque le paramètre *k* varie.

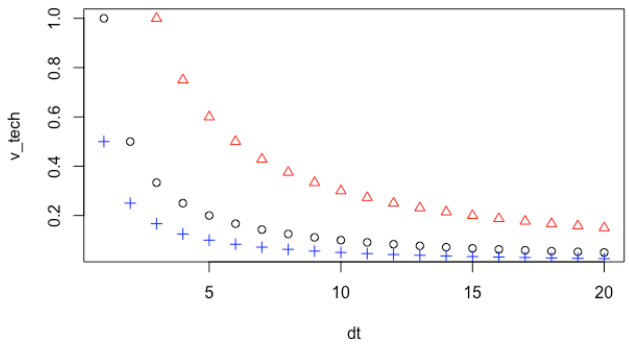

Image 1. Valeurs de v-tech pour k=0,5 (plus), k=1 (cercles), k=3 (triangles)

En appliquant cette formule à l'exemple précédent de « quadrantectomie » et « patient », les valeurs *v-tech* pour k=1 des deux termes sont respectivement : *v-tech*(quadrantectomie) = 1, *v-tech*(patient) = 0,33. Cet exemple confirme que le terme « patient » a une valeur de technicité inférieure à celle de « quadrantectomie ».

2.1 Expériences en cours

Actuellement, nous sommes en train d'expérimenter notre proposition en travaillant sur un corpus en langue française de documents médicaux concernant la vaccination. L'objectif est celui d'assigner une valeur de technicité aux termes médicaux qui ont été extraits manuellement à l'aide d'un expert en terminologie. Afin que la valeur *v-tech* puisse être assignée, il est nécessaire de disposer d'une ressource linguistique qui permette de calculer le nombre de domaines associés à une unité linguistique. Les expériences réalisées nous ont amenées à nous pencher sur le problème du manque de ressources ayant des données complètes et structurées et qui permettent une visualisation intégrale des domaines d'appartenance des termes. En fait, afin de partiellement surmonter ce problème, notre analyse est actuellement menée en rassemblant les informations fournies par trois ressources différentes : 1) le Centre National de Ressources Textuelles (CNRTL),[5] 2) le Grand Dictionnaire Terminologique (GDT)[6] et 3) Wiktionnaire.[7]

Cependant, même en s'appuyant sur trois sources, il y a des cas de termes médicaux, tels que le substantif « sévérité » associé à une maladie ou « protection » contre un virus par exemple, pour lesquels il n'existe pas

[3] http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=26527442
[4] http://www.cnrtl.fr/definition/patient
[5] http://www.cnrtl.fr
[6] http://www.granddictionnaire.com
[7] https://fr.wiktionary.org/wiki/Wiktionnaire:Page_d'accueil

une documentation complète concernant tous les domaines d'utilisation des termes. Pour cette raison, nous soutenons la nécessité de créer une ressource de données structurée qui puisse permettre l'extraction automatique des informations d'intérêt et, par conséquent, le calcul du valeur *v-tech* pour l'accomplissement de différentes tâches dans le domaine de la terminologie et de la recherche d'information.

3. Conclusions et perspectives

Dans ce résumé, nous proposons de repenser au concept de technicité des termes en tant que propriété intrinsèque et calculable sur la base des domaines auxquels ils appartiennent. Nous avons développé une définition formelle à travers la formule *v-tech* qui exprime le degré de technicité comme inversement proportionnel à la polysémie du terme. La formule proposée est générale et applicable à divers domaines et nous sommes en train de l'expérimenter sur une collection de termes médicaux.

De cette première analyse, nous avons constaté la nécessité de créer une ressource terminologique structurée qui puisse intégrer complètement les informations en ligne. De plus, la valeur *v-tech* peut être ajoutée aux formules actuelles d'*extraction automatique de termes* afin d'améliorer, en termes de *précision* et de *rappel*, la récupération de documents pertinents à un sujet, et nous nous proposons de mener des expériences dans ce sens. Enfin, spécifiquement pour le domaine médical, nous visons à approfondir la relation entre « technicité », telle que définie précédemment, et « difficulté » de compréhension dans le contexte du dialogue médecin-patient.

Références

[1] Amjadian, E., Inkpen, D., Paribakht, T. S., & Faez, F. (2018). Distributed specificity for automatic terminology extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *24*(1), 23-40.

[2] Bouamor, D., Llanos, L. C., Ligozat, A. L., Rosset, S., & Zweigenbaum, P. (2016). Transfer-Based Learning-to-Rank Assessment of Medical Term Technicality. In *LREC (Language Resources and Evaluation Conference) 2016*.

[3] Bourigault, D., Jacquemin, C., & L'Homme, M. C. (Eds.). (2001). *Recent advances in computational terminology* (Vol. 2). John Benjamins Publishing.

[4] Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, *9*(1), 99-115.

[5] Drouin, P., Grabar, N., Hamon, T., Kageura, K., & Takeuchi, K. (2018). Computational terminology and filtering of terminological information. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *24*(1), 1-6.

[7][6] Foo, J. (2012). *Computational terminology: Exploring bilingual and monolingual term extraction* (Doctoral dissertation, Linköping University Electronic Press).

[7] Grabar, N., & Hamon, T. (2016, May). A large rated lexicon with french medical words. In *LREC (Language Resources and Evaluation Conference) 2016*.

[8] Guilbert, L. (1973). La spécificité du terme scientifique et technique. *Langue française*, (17), 5-17.

[9] Humbley, J. (2016). Catherine Resche (dir.), Terminologie et domaines spécialisés, Approches plurielles. Paris: Classiques Garnier, Rencontres 143, Série linguistique 2, 2015. *ASp. la revue du GERAS*, (70), 127-132.

[10] Inkpen, D., Paribakht, T. S., Faez, F., & Amjadian, E. (2016). Term Evaluator: A Tool for Terminology Annotation and Evaluation. *Int. J. Comput. Linguistics Appl.*, *7*(2), 145-165.

[11] Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *3*(2), 259-289.

[12] Ley, P. (1988). *Communicating with patients: Improving communication, satisfaction and compliance*. Croom Helm.

[13] Planas, E. (2012). BiTermEx Un prototype d'extraction de mots composés à partir de documents comparables via la méthode compositionnelle. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*.

[14] Roche, M. (2018). Définition pluridisciplinaire de la notion de" terme". In *Proceedings of Terminology & Ontology : Theories and application conference (TOTh)*.

[15] Sandoval, A. M., Díaz, J., Llanos, L. C., & Redondo, T. (2018). Biomedical Term Extraction: NLP Techniques in Computational Medicine. *International Journal of Interactive Multimedia and Artificial Intelligence*, (In Press).

[16] Serianni, L. (2003). Italiani scritti, il Mulino.

[17] Simon, N. I., & Kešelj, V. (2018, August). Automatic Term Extraction in Technical Domain using Part-of-Speech and Common-Word Features. In *Proceedings of the ACM Symposium on Document Engineering 2018* (p. 51). ACM.

[18] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

[19] Vecchiato, S., & Gerolimich, S. (2013). La langue médicale est-elle «trop complexe»?. *Nouvelles perspectives en sciences sociales: Revue internationale de systémique complexe et d'études relationnelles*, *9*(1), 81-122.