Vers une méthodologie pour l'extraction automatique des collocations en terminologie médicale

1. Introduction

Cette étude s'inscrit dans le contexte du traitement automatique des langues (TAL) et vise à proposer une méthodologie efficace pour l'extraction automatique des collocations à collocatif verbal relevant du langage médical.

Il existe un grand nombre d'études sur les unités phraséologiques, de type locution et/ou collocation, en langue de spécialité : (Galinski, 1990) ; (Rousseau, 1993) ; (Candel, 1995) ; (Rosenbaum, 2016) pour le langage économique ; (Tutin, 2014), (Jacques et Tutin, 2018), en particulier, pour le langage scientifique. Dans ce contexte, l'axe de recherche de la « Phraséologie Computationnelle » prend forme et rassemble les intérêts de nombreux chercheurs de TAL en ce qui concerne l'écrémage, l'identification et l'extraction (semi-) automatique d'unités phraséologiques. Pour ne citer que les contributions les plus récentes, voir le chapitre de (Heid, 2008), l'enquête menée par (Constant et al. 2017) sur le traitement des expressions polylexicales, le livre rassemblant les actes du colloque Europhrase 2019[1] de (Pastor et Birukou, 2019) et, enfin, le livre récent intitulé *Computational Phraseology* et édité par (Pastor et Colson, 2020). Le processus d'extraction automatique est principalement basé sur des approches linguistiques, statistiques et/ou hybrides dont les mesures ont été importées par le domaine de la recherche d'information (RI) : *Term frequency/Inverse Document Frequency* (TF/IDF) (Salton et Yang 1973), *Information Mutuelle* (Church et Hanks 1990), *T-Score* (Church et al. 1991), *C/NC value* (Frantzi et al. 1998). Enfin, des ressources spécifiquement conçues pour cette tâche ont été réalisées afin d'augmenter les performances d'extraction, par exemple : TermoStat (Drouin, 2003), BiTermEx (Planas, 2012) et TermEvaluator (Inkpen, 2016), etc.

2. Méthodologie d'extraction automatique

L'objectif de notre étude est de proposer une méthodologie pour l'identification et l'extraction automatique des collocations à collocatif verbal dans des textes médicaux. Reprenant la définition fournie par (Polguère, 2015), nous considérons la collocation comme une structure binaire, formée d'une base et d'un collocatif, constituant une association lexicale privilégiée dont le sens est compositionnel. Notre méthodologie consistera donc, tout d'abord, à identifier la base de la collocation constituée de noms (dans le cas spécifique, les termes véhiculant une signification médicale) et, ensuite, les collocatifs verbaux associés.

Le corpus choisi comprendra tous les textes rédigés en français contenus dans le Manuel MSD disponible en ligne.[2] En perspective diastratique et multilingue, cette ressource fournit des informations spécialisées à l'usage des professionnels et des articles de vulgarisation pour consultation du grand public. Une fois sélectionné notre corpus, nous procéderons à l'étiquetage morphosyntaxique des textes en utilisant un logiciel implémenté en langage de programmation R[3] afin d'attribuer une étiquette morphosyntaxique à chaque mot contenu dans notre corpus. Après avoir rassemblé le corpus et étiqueté les textes, nous procéderons avec l'extraction automatique des termes médicaux de type nom. Pour cette dernière tâche, nous utiliserons tous

[1] http://www.lexytrad.es/europhras2019/
[2] https://www.msdmanuals.com/fr/accueil
[3] https://cran.r-project.org/web/packages/udpipe/index.html

les termes médicaux en français contenus dans le thésaurus MeSH *terms*.[4] Ce dictionnaire médical, généralement utilisé pour l'indexation d'articles biomédicaux, jouera le rôle de ressource de correspondance pour l'extraction de tous les termes médicaux dans notre corpus. Ensuite, nous mènerons une analyse de dépendance en utilisant les corpus arborés (*treebank*) fournis dans le cadre du projet *Universal Dependencies* (UD),[5] toujours à l'aide du langage de programmation R (avec le package UDpipe),[6] afin d'identifier automatiquement tous les verbes associés aux noms précédemment extraits. Cette méthodologie nous permettra, donc, d'extraire automatiquement les collocatifs verbaux associés aux bases des collocations sous la forme de noms : en considérant le terme « médicament », par exemple, nous nous attendons à extraire automatiquement tous les verbes associés au substantif afin d'en étudier les occurrences et en vérifier les associations préférentielles comme « prescrire un médicament », « administrer un médicament », « tolérer un médicament » etc.

3. Conclusions et perspectives

Dans cette étude, nous proposons une méthodologie pour l'extraction automatique des unités phraséologiques dans le langage médical. En particulier, nous nous concentrerons sur l'identification des collocations à collocatif verbal. À partir de cette analyse et compte tenu de la variété diastratique du lexique médical, c'est-à-dire de la terminologie utilisée entre les professionnels de la santé et de la terminologie destinée aux profanes (Vecchiato et Gerolimich, 2013), nous visons à mener une étude contrastive future afin de vérifier s'il existe des différences dans la représentation du phénomène de la collocation dans les deux variations de registre.

Références

Candel, D. (1995). Locutions en langues de spécialité. *Cahiers du français contemporain*, *2*, 151-173.

Church, Kenneth Ward, and Patrick Hanks. (1990). "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1): 22–29.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle (1991) "Using Statistics in Lexical Analysis." *Lexical Acquisition: Exploiting on-Line Resources to Build a Lexicon* 115: 164.

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 837-892.

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

Frantzi, Katerina T, Sophia Ananiadou, and Junichi Tsujii. (1998) "The C-Value/Nc-Value Method of Automatic Recognition for Multi-Word Terms." In *International Conference on Theory and Practice of Digital Libraries*, 585–604. Springer.

[4] https://meshb.nlm.nih.gov/search
[5] https://universaldependencies.org
[6] https://cran.r-project.org/web/packages/udpipe/index.html

Galinski, C. (1990). Terminology and phraseology. Terminology Science and Research. En Journal of the International Institute for Terminological Research (IITF), 1.

Jacques, M. P., & Tutin, A. (2018). *Lexique transversal et formules discursives des sciences humaines*. ISTE Group.

Heid, U. (2008). Computational phraseology. An overview. *Phraseology: An Interdisciplinary Perspective; Granger, S., Meunier, F., Eds*, 337-360.

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.

Pastor, C., & Birukou. (2019). Computational and Corpus-Based Phraseology. Springer International Publishing.

Pastor, G. C., & Colson, J. P. (2020). Computational Phraseology. John Benjamins Publishing Company.

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*.

Polguère, Alain (2015), « Non compositionnalité : ce sont toujours les locutions faibles qui trinquent », dans Verbum, n° 37(2), pp. 257-280.

Rosenbaum Franková, L. (2016). Phrasèmes spécialisés dans les textes économiques. Cahiers de lexicologie, 2016(108), 43-57.

Rousseau, L. J. (1993). Terminologie et phraséologie, deux composantes indissociables des langues de spécialités. *Terminologies nouvelles–Phraséologie Actes du séminaire international*, (10), 9-11.

Salton, Gerard, and Chung-Shu Yang, (1973) *"On the Specification of Term Values in Automatic Indexing."* Journal of Documentation 29 (4): 351–372.

Tutin, A. (2014). *L'écrit scientifique : du lexique au discours*. F. Grossmann, & P. U. de Rennes (Eds.). Presses universitaires de Rennes.

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.