Traitement Automatique de Langues pour la simplification de documents de santé

Natalia Grabar, Anaïs Koptient, Rémi Cardon

AfIA
Association française
pour l'Intelligence Artificielle

Traitement Automatique de Langues pour la simplification de documents de santé

CNRS, UMR 8163 STL, Université de Lille
http://natalia.grabar.free.fr

Natalia GRABAR
natalia.grabar@univ-lille.fr

Anaïs KOPTIENT
an.koptient@hotmail.fr

Rémi CARDON
remi.cardon@univ-lille.fr

Introduction

La simplification automatique vise à fournir une version simplifiée des textes. La simplification peut concerner le lexique, la syntaxe, la sémantique mais aussi la pragmatique et l'organisation des textes. La simplification peut être vue comme une aide fournie aux lecteurs [19, 6] ou comme un pré-traitement pour d'autres applications de TAL [5, 21].

Nous nous intéressons à la simplification de documents médicaux car ils comportent une terminologie technique (*myocarde, cholécystectomie, aplasie*). De tels termes sont souvent incompréhensibles par les patients. Le travail décrit est effectué dans le cadre du projet *CLEAR (Communication, Literacy, Education, Accessibility, Readability)* financé par l'ANR sous référence ANR-17-CE19-0016-01.

Travaux existants en simplification

Ces dernières années, la simplification automatique a connu un grand succès auprès des chercheurs. Il existe actuellement trois types de méthodes qui sont exploitées pour effectuer la simplification :

- *Méthodes guidées par les connaissances et les règles*. Ces méthodes exploitent des ressources de type WordNet [4], les fréquences [8, 7] ou la longueur [1] des mots ;
- *Méthodes issues de la traduction automatique*, où la simplification correspond à la traduction monolingue allant d'un texte technique vers un texte simple [22, 18] ;
- *Méthodes exploitant exploitant la sémantique distributionnelle*. Entraînés sur des données convenables, les vecteurs de mots peuvent contenir des équivalents plus simples de termes, qui peuvent alors être exploités pour effectuer la simplification [9, 13].

La majorité des travaux existants sont effectués sur les documents en anglais et exploitent les données provenant de Wikipedia et Simple Wikipedia ou de Newsela.

Nos contributions

Nous avons proposé plusieurs contributions, qu'elles soient méthodologiques ou de ressources, selon les différents aspects de la simplification automatique :

1. création d'un corpus avec des documents comparables en français, qui met en face les documents médicaux en version technique avec leur version simplifiée ou simple [10]. Ce corpus regroupe les informations sur les médicaments (RCP et notices de médicaments), les résumés des revues systématiques Cochrane et des articles encyclopédiques (Wikipedia et Vikidia) ;

2. méthodes pour la recherche de phrases parallèles au sein de documents comparables [2]. Ces méthodes effectuent la catégorisation des paires de phrases en deux catégories : phrases alignables ou non alignables. Plusieurs descripteurs sont exploités, comme les descripteurs lexicaux (nombre de mots communs, longueur des phrases), les descripteurs basés sur les chaînes d'édition, les descripteurs basés sur les similarités lexicales (cosinus, Dice et Jaccard), les descripteurs basés sur les n-grammes de mots et de caractères, et les descripteurs basés sur les plongements lexicaux. Ces travaux ont permis de constituer un ensemble de plus de 10 000 couples de phrases alignées ;
3. annotation de couples de phases alignées selon les transformations dues à la simplification et élaboration d'une typologie des procédés de la simplification [14] ;
4. méthodes pour la détection de mots et passages difficiles dans les documents médicaux : (1) classification de mots grâce à l'exploitation d'un lexique médical annoté en difficulté [12], (2) exploitation de méthodes d'oculométrie, où les fixations plus longues et les saccades plus courtes indiquent la présences de passages plus difficiles à comprendre [11], (3) exploitation de méthodes neuronales avec le lexique médical annoté en difficulté [20] ;
5. création d'un lexique où les termes médicaux techniques sont associés avec leurs paraphrases grand public, comme {*myocarde* ; *muscle du cœur*} ou {*cholécystectomie* ; *ablation de la vésicule biliaire*}. Ce lexique contient actuellement 7 580 paraphrases pour 4 516 termes médicaux. De plus, les termes et les paraphrases du lexique reçoivent des indices de leur lisibilité [16]. À terme, ce lexique sera aligné avec les concepts d'UMLS [17] afin de rendre cette ressource plus facilement exploitable ;
6. méthodes pour la simplification automatique de documents médicaux : une méthode à base de règles [15] et une méthode exploitant les approches neuronales issues de la traduction automatique [3]. Ce dernier travail a permis également de produire la traduction française du corpus WikiLarge [23] avec 300 000 couples de phrases.

Ces différentes contributions ont permis de faire des avancées sur différents aspects liés à la simplification automatique. Notons que plusieurs des ressources constituées sont librement disponibles pour la recherche [1]. Nous espérons que ces ressources pourront motiver les travaux autour de la simplification en français.

Références

[1] Susana Bautista, Pablo Gervás, and R. Ignacio Madrid. Feasibility analysis for semi-automatic conversion of text to improve readability. In *Int Conf on Inform and Comm Technology and Accessibility (ICTA)*, pages 33–40, 2009.

[2] Rémi Cardon and Natalia Grabar. Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français. volume 61, pages 15–39, 2020.

[3] Rémi Cardon and Natalia Grabar. French biomedical text simplification : When small and precise helps. In *COLING 2020*, pages 1–8, 2020.

[4] J Carroll, G Minnen, Y Canning, S Devlin, and J Tait. Practical simplification of English newspaper text to assist aphasic readers. In *AAAI-98 Workshop on Integrating Artifficial Intelligence and Assistive Technology*, pages 7–10, 1998.

1. http ://natalia.grabar.free.fr/resources.php

[5] R Chandrasekar and B Srinivas. Automatic induction of rules for text simplification. *Knowledge Based Systems*, 10(3) :183–190, 1997.

[6] Ping Chen, John Rochford, David N. Kennedy, Soussan Djamasbi, Peter Fay, and Will Scott. Automatic text simplification for people with intellectual disabilities. In *AIST*, pages 1–9, 2016.

[7] Jan De Belder and Marie-Francine Moens. Text simplification for children. In *Workshop on Accessible Search Systems of SIGIR*, pages 1–8, 2010.

[8] Siobhan Devlin and John Tait. The use of psycholinguistic database in the simplification of text for aphasic readers. In *Linguistic Database*, pages 161–173, 1998.

[9] Goran Glavas and Sanja Stajner. Simplifying lexical simplification : Do we need simplified corpora ? In *ACL-COLING*, pages 63–68, 2015.

[10] Natalia Grabar and Rémi Cardon. Clear – simple corpus for medical French. In *Workshop on Automatic Text Adaption (ATA)*, pages 1–11, 2018.

[11] Natalia Grabar, Emmanuel Farce, and Laurent Sparrow. Study of readability of health documents with eye-tracking approaches. In *Workshop on Automatic Text Adaption (ATA)*, pages 1–11, 2018.

[12] Natalia Grabar and Thierry Hamon. A large rated lexicon with French medical words. In *LREC (Language Resources and Evaluation Conference)*, pages 1–12, 2016.

[13] Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. SimpleScience : Lexical simplification of scientific terminology. In *EMNLP*, pages 1–6, 2016.

[14] Anaïs Koptient, Rémi Cardon, and Natalia Grabar. Simplification-induced transformations : typology and some characteristics. In *Proc of the 18th BioNLP Workshop and Shared Task*, pages 309–318, Florence, Italy, 2019.

[15] Anaïs Koptient and Natalia Grabar. Fine-grained text simplification in french : steps towards a better grammaticality. In *Proc of ISHIMR 2020*, pages 1–10, 2020.

[16] Anaïs Koptient and Natalia Grabar. Rated lexicon for the simplification of medical texts. In *Proc of HEALTHINFO 2020*, pages 1–6, 2020.

[17] DA Lindberg, BL Humphreys, and AT McCray. The Unified Medical Language System. *Methods Inf Med*, 32(4) :281–291, 1993.

[18] Sergiu Nisioi, Sanja Stajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring neural text simplification models. In *Ann Meeting of the Assoc for Comp Linguistics*, pages 85–91, 2017.

[19] Gustavo H. Paetzold and Lucia Specia. Benchmarking lexical simplification systems. In *LREC*, pages 3074–3080, 2016.

[20] Hanna Pylieva, Artem Chernodub, Natalia Grabar, and Thierry Hamon. Generalizability of readability models for medical terms. In *MEDINFO 2019*, pages 1–5, 2019.

[21] Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. SimConcept : A hybrid approach for simplifying composite named entities in biomedicine. In *BCB '14*, pages 138–146, 2014.

[22] S Wubben, A van den Bosch, and E Krahmer. Sentence simplification by monolingual machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1015–1024, 2012.

[23] Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing*, pages 584–594, 2017.