Utilisation d'une base de connaissances de spécialité et de sens commun pour la simplification de comptes-rendus radiologiques

Lionel Ramadier[1] Mathieu Lafourcade[2]
(1) LIMSI, Campus universitaire bât 508 Rue John von Neumann, 91405 Orsay, France
(2) LIRMM, 161 rue Ada, 34095 Montpellier, France
ramadier@limsi.fr, mathieu.lafourcade@lirmm.fr

RÉSUMÉ

Dans le domaine médical, la simplification des textes est à la fois une tâche souhaitable pour les patients et scientifiquement stimulante pour le domaine du traitement automatique du langage naturel. En effet, les comptes rendus médicaux peuvent être difficile à comprendre pour les non spécialistes, essentiellement à cause de termes médicaux spécifiques (*prurit*, par exemple). La substitution de ces termes par des mots du langage courant peut aider le patient à une meilleure compréhension. Dans cet article, nous présentons une méthode de simplification dans le domaine médical (en français) basée sur un réseau lexico-sémantique. Nous traitons cette difficulté sémantique par le remplacement du terme médical difficile par un synonyme ou terme qui lui est lié sémantiquement à l'aide d'un réseau lexico-sémantique français. Nous présentons dans ce papier, une telle méthode ainsi que son évaluation.

ABSTRACT

Radiological text simplification using a general knowledge base.

In the medical domain, text simplification is both a desirable and a challenging natural language processing task. Indeed, first, medical texts can be difficult to understand for patient, because of the presence of specialized medical terms. Replacing these difficult terms with easier words can lead to improve patient's understanding. In this paper, we present a lexical network based method to simplify health information in French language. We deal with semantic difficulty by replacement difficult term with supposedly easier synonyms or by using semantically related term with the help of a French lexical semantic network. We extract semantic and lexical information present in the network. In this paper, we present such a method for text simplification along with its qualitative evaluation. .

MOTS-CLÉS : TALN, médical, simplification, réseau lexico-sémantique.

KEYWORDS: NLP, BioNLP, text simplification, lexico-semantic network.

1 Introduction

La simplification de texte (ST) est un domaine du traitement automatique du langage naturel (TALN) dont le but est de rendre des textes plus compréhensibles tout en garantissant l'intégrité de leur contenu et de leur structure. Dès lors, la ST peut être un moyen d'aider des personnes à accéder à la compréhension de documents écrits spécialisés. En effet, les problèmes de compréhension sont souvent dus à une grande complexité des textes, tant au niveau lexical que syntaxique. Dans ce cadre, la ST peut être vue comme une tâche de traduction mono-langue, où la langue source a besoin d'être

traduite en une version simplifiée de la même langue.

Son application dans le domaine médical revêt une importance particulière. La compréhension d'un texte médical peut être particulièrement ardue pour les patients non spécialistes du domaine médical. Les textes médicaux sont difficiles à comprendre pour un non expert (Keselman & Smith, 2012) du fait que les médecins écrivent souvent avec des termes spécialisés (*ataxie*) et des abréviations (*SA* pour *sans aménorrhée*) qui nécessitent une certaine connaissance médicale. (Chapman *et al.*, 2003) et (Lerner *et al.*, 2000) ont déjà montré que les termes médicaux pouvaient être un obstacle à la compréhension pour les patients. Ces difficultés peuvent avoir un impact négatif sur la communication entre les patients et les médecins, et les soins offerts aux malades (Tran *et al.*, 2009).

Cependant, il existe peu de travaux sur des méthodes automatiques de simplification des textes médicaux et plus particulièrement des comptes rendus cliniques (Keselman *et al.*, 2008). Le but de notre travail est de simplifier des comptes rendus radiologiques en français grâce à une base de connaissance de sens commun qui contient à la fois de la connaissance générale mais aussi de spécialité. Notre approche concerne la simplification lexicale. Nous utilisons pour cela non seulement des synonymes mais aussi des termes liés par des relations hiérarchiques et/ou sémantiques. Nous présentons dans un premier temps les travaux liés à l'état de l'art (section 2). Nous présentons ensuite notre approche de simplification utilisée (section 3). Nous décrivons et discutons les résultats obtenus (sections 4 et 5). Nous concluons avec des perspectives pour les travaux futurs.

2 Etat de l'art

Le niveau de difficulté peut varier entre différents types de textes médicaux (Leroy *et al.*, 2010) et même les brochures pour patient peuvent être difficile à comprendre (Kokkinakis *et al.*, 2012). La simplification lexicale aide à rendre un texte plus compréhensible. En effet, (Abrahamsson *et al.*, 2014) ont montré que le remplacement des mots difficiles par des synonymes du langage courant pouvait réduire le niveau de difficulté d'un texte. La substitution par les synonymes a été évaluée sur des textes médicaux anglais (Leroy *et al.*, 2012) (Slaughter *et al.*, 2005) mais aussi suédois (Keskisärkkä, 2012). (Leroy *et al.*, 2012) utilise un lexique de synonymes et remplace les termes difficiles avec des synonymes du langage courant. Le niveau de difficulté d'un mot est déterminé par sa fréquence d'apparition dans un corpus général. Dans (Zeng-Treitler *et al.*, 2007), les auteurs utilisent deux stratégies pour réduire la difficulté lexicale des textes médicaux (le remplacement par des synonymes et par l'ajout d'explication). Pour effectuer la simplification des textes à l'aide de synonymes par exemple, des ressources spécifiques sont nécessaires. Dans le domaine médical, ces ressources se présentent souvent sous forme de lexiques où les termes sont mis en correspondance avec les expressions non spécialisées correspondantes. L'utilisation de ce type de lexique est apparue avec le travail collaboratif Consumer Health Vocabulary (CHV) (Zeng-Treitler *et al.*, 2007) (Qenam *et al.*, 2017). (Elhadad & Sutaria, 2007) a construit un lexique de termes alignés avec leurs équivalents non techniques à partir de l'UMLS (Unified Medical Language System [1]). (Leroy *et al.*, 2012) ont développé un système qui utilise la familiarité d'un terme pour identifier la difficulté du texte et sélectionne des termes plus compréhensibles à partir de ressources lexicales comme WordNet [2], UMLS et Wiktionnary [3]. Dans le domaine de la radiologie, plusieurs études ont montré que les comptes rendus radiologiques sont parmi les plus difficiles à comprendre (Keselman *et al.*, 2007). Une

1. https ://www.nlm.nih.gov/research/umls/
2. https ://wordnet.princeton.edu/
3. https ://www.wiktionary.org/

équipe suédoise (Kvist & Velupillai, 2013) a développé un corpus de comptes-rendus radiologiques pouvant être utilisé pour le développement d'outils de simplification de textes médicaux.

Une autre approche consiste à détecter les composants morphologiques des mots difficiles. Il peut être intéressant de décomposer un terme comme *hématurie* en ses composants *hématie, urine*. Cette décomposition automatique des termes se base souvent sur des méthodes à base de règles ou des approches probabilistes en corpus (Namer, 2003), (Claveau & Kijak, 2014). (Grabar & Hamon, 2015) ont développé un système qui permet d'acquérir des paraphrases non spécialisées pour des termes techniques composés du domaine médical.

3 Notre approche

Dans ce papier, nous étudions la ST d'un type de document médical à savoir les comptes rendus de radiologie. Nous utilisons une méthode de remplacement lexicale non seulement par des synonymes mais aussi par d'autres relations (par exemple, l'hyperonymie). Cette dernière peut être très utile car un terme peut être expliqué comme une incidence spécifique de ses parents. Par exemple, une *carcinome hépatocellulaire* est un *cancer du foie* (relation *is-a*). La base de connaissance sur laquelle se base notre méthode de simplification est le réseau lexico-sémantique JeuxDeMots [4] (JDM) (Lafourcade, 2007).

3.1 Ressources

Deux ressources sont utilisées dans notre approches : (1) une base de connaissances et (2) un corpus de comptes-rendus radiologiques.

Le réseau lexical JeuxDeMots (JDM)

Le réseau JDM est un graphe lexico-sémantique pour la langue française dont les relations entre les termes sont capturées par la combinaison d'un Game With A Purpose (GWAP) (Lafourcade, 2007) avec un outil contributif nommé Diko (contribution manuelle et inférences automatiques avec validation (Zarrouk *et al.*, 2013)). Le réseau contient à la fois des connaissances du domaine général ainsi que des connaissances spécialisées. La partie médicale de la ressource JeuxDeMots a été constituée à partir d'un corpus de comptes rendus cliniques, des pages médicales de wikipédia, du site des maladies rares Orphanet [5] et du Dictionnaire médical de l'Académie de Médecine [6]. En février 2018, le réseau JDM contient 181803113 relations et 2712500 termes. Le tableau suivant (table 1) donne un ordre de grandeur de la quantité d'information que nous avons à notre disposition dans le domaine de la radiologie.

Termes	liens sortants	liens entrants
maladie	16648	21040
anatomie	50 846	94 000
radiologie	733	940

TABLE 1 – Nombre de relations de certains mots clefs dans le réseau JDM.

4. http ://www.jeuxdemots.org/
5. http ://www.orpha.net/consor/cgi-bin/index.php ?lng=FR
6. http ://dictionnaire.academie-medecine.fr/

r_isa termes génériques.
r_synonym synonymes or quasi-synonymes.
r_syn_strict synonymes stricts.
r_equiv acronyme ou abréviation.

FIGURE 1 – Les relations utilisées pour la simplification de textes médicaux

Il existe 80 types de relations dans le réseau mais dans ce travail nous utilisons seulement 4 types différents de relations (figure 1).

Nous utilisons ce réseau car les relations entre les termes sont à la fois pondérées et annotées (Ramadier *et al.*, 2014). Dans le réseau JDM, les relations sont pondérées c'est à dire que le poids reflète la force d'association entre les termes. Par contre, dans le domaine des connaissances spécialisées, la corrélation entre la force d'association de la relation et son importance conceptuelle n'est pas toujours assurée. C'est pourquoi, il est apparu intéressant d'utiliser des annotations pour certaines relations. Ces annotations peuvent nous aider dans la tâche de simplification lexicale (grâce à l'annotation *langage courant*). L'implémentation d'une annotation se fait par réification de la relation à annoter dans le réseau lexical. Le nœud relation créé peut être associé à d'autres termes. L'annotation de relation n'est qu'un type de relation parmi d'autres. Les valeurs d'annotation sont des termes standards (*fréquent, rare, langage courant, etc.*).

Le corpus de comptes-rendus radiologiques

Notre corpus est constitué de 35 000 comptes rendus radiologiques représentant différentes modalités d'imagerie médicale (imagerie par résonance magnétique, scanner, radiographie, échographie, radiologie interventionnelle). Ces comptes rendus sont écrits de façon semi-structurés. Ils sont, en général, divisés en quatre parties, chacune étant écrite de manière non structurée (avec souvent quantité d'acronymes et d'abréviations). Ils ont été, au préalable, dé-identifiés.

3.2 Méthode

Dans un premier temps, il est important d'identifier les mots qui posent le plus de difficulté de compréhension aux patients (Kauchak *et al.*, 2017). Nous sélectionnons les termes difficiles en utilisant une méthode classique, à savoir la fréquence des termes (TF) et la fréquence documentaire (DF) pour calculer l'IDF (Inverse Document Frequency). La reconnaissance des termes composés est effectuée en amont par comparaison au contenu de JDM. Si un terme est polysémique, nous sélectionnons le terme en lien avec la médecine (par exemple, pour kyste nous choisissons le raffinement médecine (*kyste >médecine*) et non celui liés à la botanique (*kyste>botanique*)). Pour choisir un terme du langage courant, nous utilisons les annotations de relations (Ramadier *et al.*, 2014). Si la relation a une annotation *langage courant*, alors le terme médical original est remplacé par son synonyme ou son hyperonyme. Nous présentons un exemple de simplification à la figure 2. Nous remplaçons systématiquement le terme *antérieur* par *en avant* et *postérieur* par *derrière* . Certaines abréviations (*AVP*) seront remplacées par leur signification (*accident de la voie publique*).

Si un terme composé difficile à comprendre n'a pas d'annotation de relations, nous extrayons l'information sémantique des mots qui le composent. En effet, dans le réseau l'information lexicale indique si un mot fait partie du langage courant ou pas.

(a) le patient a une *aphasie* depuis deux jours. → Le patient a un **mutisme** depuis deux jours.
(b) Le patient se plaint de *céphalée* → Le patient se plaint de **maux de tête**
(c) Symptôme : *hématurie* → Symptôme : **sang dans les urines**

FIGURE 2 – Exemples de simplification de terme. Le terme remplacé est en gras.

4 Expérience et résultats

4.1 Expérience

Nous utilisons un sous échantillon de notre corpus (200 comptes rendus) et nous les simplifions grâce à notre approche. Pour l'évaluation manuelle, 250 phrases ont été sélectionnées aléatoirement pour une évaluation humaine manuelle ainsi que pour le test d'évaluation standard pour la compréhension (*cloze test* ou *texte à trou*) (Taylor, 1953). Selon la procédure standard du test, le 5ème mot de chaque phrase est remplacé par un espace blanc. Nous avons recruté 4 personnes (non expertes du domaine médical) pour réaliser le test. Chaque personne réalisait le test sur le compte rendu original et sa version simplifiée. Un expert vérifiait, ensuite, manuellement les réponses pour évaluer leur exactitude. Nous calculions, alors un score (*cloze score* (Zeng-Treitler *et al.*, 2007)) qui représentait le pourcentage de réponses exactes correspondant aux mots effacés.

4.2 Résultats

En moyenne, 10 termes furent simplifiés par compte rendu. La plupart des simplifications (75%) ont été déclarés correctes par l'évaluateur humain. Pour 12% des phrases, le mot remplacé avait un sens légèrement différent par rapport au terme original. Ces erreurs peuvent s'expliquer par le fait que dans certains cas le synonyme n'est pas strict. Par exemple, *kyste* et *abcès* sont synonymes ou quasi-synonymes dans le réseau mais dans le domaine médical, leur sens sont différents. Un autre type d'erreur est que parfois pour la relation d'hyperonymie le mot remplacé était trop général (par exemple, *appendagite* a été remplacé par *maladie*). Nous montrons quelques exemples de termes originaux et leurs synonymes du langage courant (Table 2).

Terme original	**Remplacé par**
aphasie	mutisme
prurit	démangeaison
dyspnée	difficulté à respirer
glioblastome	tumeur maligne du cerveau
arthrite	inflammation des articulations

TABLE 2 – Exemples de termes simplifiés.

La table 3 et table 4 montrent les résultats pour les comptes rendus originaux et simplifiés. Le tableau 3

montre les résultats obtenus avec seulement les annotations de relations (*langage courant*). Le tableau 4 indique les résultats en utilisant à la fois les annotations mais aussi les informations sémantiques présentes dans le réseau JDM.

Comptes rendus originaux	Comptes rendus simplifiés
18%	48%

TABLE 3 – *Cloze score* pour les comptes rendus originaux et simplifiés avec seulement les annotations.

Terme original	Comptes rendus simplifiés
18%	57%

TABLE 4 – *Cloze score* pour les comptes rendus originaux et simplifiés quand on utilise non seulement les annotations mais aussi les informations sémantiques présentes dans le réseau JDM

Le score de 18% sur les comptes rendus originaux prouve que les comptes rendus radiologiques sont difficilement compréhensibles pour les patients.

5 Discussion

Nous décrivons un système de simplification pour un corpus français de comptes rendus radiologiques. Le score (cloze score) pour les textes originaux sont plus faibles que dans d'autres études (Zeng-Treitler *et al.*, 2007) qui traitent des textes médicaux variés (comptes rendus chirurgicaux, lettres de sortie). Les comptes rendus radiologiques sont parmi les plus difficiles à comprendre pour des non-experts. Nous avons implémenté un système basé une simplification lexicale dont le but est d'offrir une meilleure compréhension aux patients. Notre méthode se base sur le réseau JDM et en particulier, sur les annotations de relations pour choisir un terme plus compréhensible. 80% des termes remplacés sont utiles. Si nous utilisons seulement les annotations de relations pour la tâche de simplification, nous obtenons un score de 48%. En utilisant en complément, l'information lexicale présente dans le réseau, nous améliorons notre résultat et nous atteignons un score de 57%. Mais 35% des termes difficiles qui auraient du être remplacés ne l'ont pas été parce qu'il n'y avait pas d'annotations (*langage courant*) dans le réseau. L'évaluation manuelle a aussi montré que le sens original a parfois été légèrement modifié dans certaines phrases. Dans certains cas, les mots ne sont pas strictement synonymes (*œdème* et *gonflement*, par exemple). Nos résultats sont proches de ceux de (Zeng-Treitler *et al.*, 2007) bien que notre corpus soit plus important et ne contienne que des comptes-rendus radiologiques. Notre système est évidement perfectible. Une piste d'amélioration possible est d'augmenter la couverture des annotations au sein du réseau. Nous avons également l'intention de simplifier la syntaxe des phrases dans un deuxième temps.

6 Conclusion

Nous avons développé un système dont l'objectif est d'améliorer la compréhension des comptes-rendus radiologiques par le patient. Les résultats présentés ici restent préliminaires mais sont néanmoins très prometteurs. Dans ce travail, nous avons utilisé le réseau lexico-sémantique JeuxDeMots

comme base de connaissance. Bien que ce réseau soit général, il contient de nombreuses données spécialisées, notamment en médecine et radiologie pouvant être utiles dans le cadre de la tâche de simplification pour ces domaines.

Dans un travail futur, une autre tâche est de simplifier la syntaxe des textes médicaux. Une étude précédente (Campbell & Johnson, 2001) a montré des différences significatives du contenu syntaxique et la complexité entre les rapports de sortie d'hospitalisation et l'anglais quotidien. Une autre étude a souligné la difficulté de simplification syntaxique de textes (Kandula *et al.*, 2010). Pour cette tâche, nous pourrions réaliser une simplification de grammaire (par exemple, les longues phrases pourraient être découpées en phrases plus courtes). Nous planifions aussi de tester notre approche dans d'autres domaines médicaux, comme par exemple l'oncologie, parce que JDM contient des données de ce domaine.

Références

ABRAHAMSSON E., FORNI T., SKEPPSTEDT M. & KVIST M. (2014). Medical text simplification using synonym replacement : Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, p. 57–65.

CAMPBELL D. A. & JOHNSON S. B. (2001). Comparing syntactic complexity in medical and non-medical corpora. In *Proceedings of the AMIA Symposium*, p.9̃0 : American Medical Informatics Association.

CHAPMAN K., ABRAHAM C., JENKINS V. & FALLOWFIELD L. (2003). Lay understanding of terms used in cancer consultations. *Psycho-Oncology*, **12**(6), 557–566.

CLAVEAU V. & KIJAK E. (2014). Generating and using probabilistic morphological resources for the biomedical domain. In *9th edition of the Language Resources and Evaluation Conference, LREC 2014*, p. 7–p.

G. DIAS, Ed. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

ELHADAD N. & SUTARIA K. (2007). Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007 : Biological, Translational, and Clinical Language Processing*, p. 49–56 : Association for Computational Linguistics.

GRABAR N. & HAMON T. (2015). Extraction automatique de paraphrases grand public pour les termes médicaux. *Actes de TALN*.

KANDULA S., CURTIS D. & ZENG-TREITLER Q. (2010). A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, volume 2010, p. 366 : American Medical Informatics Association.

KAUCHAK D., LEROY G. & HOGUE A. (2017). Measuring text difficulty using parse-tree frequency. *Journal of the Association for Information Science and Technology*, **68**(9), 2088–2100.

KESELMAN A., LOGAN R., SMITH C. A., LEROY G. & ZENG-TREITLER Q. (2008). Developing informatics tools and strategies for consumer-centered health communication. *Journal of the American Medical Informatics Association*, **15**(4), 473–483.

KESELMAN A., SLAUGHTER L., ARNOTT-SMITH C., KIM H., DIVITA G., BROWNE A., TSAI C. & ZENG-TREITLER Q. (2007). Towards consumer-friendly phrs : patients' experience with

reviewing their health records. In *AMIA Annual Symposium Proceedings*, volume 2007, p. 399 : American Medical Informatics Association.

KESELMAN A. & SMITH C. A. (2012). A classification of errors in lay comprehension of medical documents. *Journal of biomedical informatics*, **45**(6), 1151–1163.

KESKISÄRKKÄ R. (2012). Automatic text simplification via synonym replacement.

KOKKINAKIS D., FORSBERG M., KOKKINAKIS S. J., SMITH F. & ÖHLEN J. (2012). Literacy demands and information to cancer patients. In *International Conference on Text, Speech and Dialogue*, p. 64–71 : Springer.

KVIST M. & VELUPILLAI S. (2013). Professional language in swedish radiology reports–characterization for patient-adapted text simplification. In *Scandinavian Conference on Health Informatics 2013, Copenhagen, Denmark, August 20, 2013*, p. 55–59 : Linköping University Electronic Press.

LAFOURCADE M. (2007). Making people play for lexical acquisition with the jeuxdemots prototype. In *SNLP'07 : 7th international symposium on natural language processing*, p.7̃.

LERNER E. B., JEHLE D. V., JANICKE D. M. & MOSCATI R. M. (2000). Medical communication : do our patients understand ? *The American journal of emergency medicine*, **18**(7), 764–766.

LEROY G., ENDICOTT J. E., MOURADI O., KAUCHAK D. & JUST M. L. (2012). Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In *AMIA Annual Symposium Proceedings*, volume 2012, p. 522 : American Medical Informatics Association.

LEROY G., HELMREICH S. & COWIE J. R. (2010). The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, **79**(6), 438–449.

NAMER F. (2003). Automatiser l'analyse morpho-sémantique non affixale : le système dérif. *Cahiers de grammaire*, **28**, 31–48.

QENAM B., KIM T. Y., CARROLL M. J. & HOGARTH M. (2017). Text simplification using consumer health vocabulary to generate patient-centered radiology reporting : Translation and evaluation. *Journal of medical Internet research*, **19**(12).

RAMADIER L., ZARROUK M., LAFOURCADE M. & MICHEAU A. (2014). Inferring relations and annotations in semantic network : Application to radiology. *Computación y Sistemas*, **18**(3), 455–466.

SLAUGHTER L., KESELMAN A., KUSHNIRUK A. & PATEL V. L. (2005). A framework for capturing the interactions between laypersons' understanding of disease, information gathering behaviors, and actions taken during an epidemic. *Journal of biomedical informatics*, **38**(4), 298–313.

TAYLOR W. L. (1953). "cloze procedure" : A new tool for measuring readability. *Journalism Bulletin*, **30**(4), 415–433.

TRAN T. M., CHEKROUD H., THIERY P. & JULIENNE A. (2009). Internet et soins : un tiers invisible dans la relation médecine/patient. *Ethica Clinica*, **53**, 34–43.

ZARROUK M., LAFOURCADE M. & JOUBERT A. (2013). Inference and reconciliation in a crowdsourced lexical-semantic network. *Computación y Sistemas*, **17**(2).

ZENG-TREITLER Q., GORYACHEV S., KIM H., KESELMAN A. & ROSENDALE D. (2007). Making texts in electronic health records comprehensible to consumers : a prototype translator. In *AMIA Annual Symposium Proceedings*, volume 2007, p. 846 : American Medical Informatics Association.