Fine-grained Simplification of Medical Documents

Anaïs KOPTIENT[a,1] and Muriel LONDRES[b] and Natalia GRABAR[a,]
[a] *CNRS, Univ Lille, UMR 8163 – STL, F-59000 Lille, France*
[b] *LEPS, Université Sorbonne Paris Nord, France*

Abstract. Easy access to medical and health information for children, foreigners and patients is an important issue for the modern society and research. Indeed, misunderstanding of medical and health information by patients may have a negative impact on their healthcare process and health. Even if several simplification guidelines exist, they are difficult to use by medical experts (i.e. lack of time, difficulty to respect the criteria). Existing simplification systems mainly address some lexical or syntactic transformations. We propose to combine lexical and syntactic simplifications within one rule-based system and to make the process fine-grained thanks to a better control of the grammaticality of sentences.

Keywords. Simplification, Grammaticality, Evaluation, French

1. Introduction

The issue on easy access to medical and health information by human users (*i.e.* children, foreigners, patients) is an important question for the modern society and research. Indeed, the misunderstanding of medical and health information by patients may negatively impact their healthcare process. Several simplification guidelines [1,2,3] have been proposed. They indicate what should be modified in order to simplify texts. Yet, manual simplification is a long and tedious process. Automatic simplification is a recent research topic. It is usually done using three kinds of methods: distributional probabilities, such as word embeddings [4,5], automatic translation [6,7,8,9] and rule-based [10,11,12]. We propose a rule-based method, which permits to have a better control over the grammaticality of the results. Existing simplification guidelines provide some linguistic phenomena to be taken into account, among which: use short and frequent words, avoid abbreviations, make sentences syntactically simple, avoid passive sentences, use personal style, and explain difficult concepts. We work with medical texts in French. In what follows, we present our rule-based system for text simplification in French that we evaluate along three criteria (grammaticality, simplicity and semantics). We then present the results obtained and their evaluation. Finally, we conclude and outline future work.

[1] Corresponding Author, Corresponding author, Book Department, IOS Press, Nieuwe Hemweg 6B, 1013 BG Amsterdam, The Netherlands; E-mail: bookproduction@iospress.nl.

2. Methods

Our methods cover lexical (term substitution and abbreviation explanation) and syntactic (sentence division and passive to active sentence transformation) simplification. During these transformations, we control the grammaticality of simplified sentences. We also present how the evaluation is performed. The simplification starts with plain texts, which are syntactically analyzed by the Cordial parser [13] and split into syntactic groups.

2.1. Lexical Simplification

For lexical substitution, we use a previously created lexicon [14], which currently contains 16,787 entries. Each entry is composed of a medical term and its simpler equivalent: {*abcès (abscess)*, *accumulation de pus (collection of pus)*}. Hence, for each syntactic group, we look if this group is the candidate for simplification. If so, the medical term is replaced by its simpler equivalent. Then, we verify if determiners, adjectives, or past participles of the sentence should be concorded with the new syntactic group. In the example below, the term *syncope*, which is a feminine noun, is substituted by its simpler equivalent *évanouissement ou sensation d'évanouissement* (*fainting or a faint feeling*), which syntactic head is a masculine noun. Therefore, the feminine determiner *une (a)* must be transformed in a masculine determiner *un (a)*:

- source: *Elle peut conduire très rarement à une syncope (It may rarely cause the syncope).*
- simplified: *Elle peut conduire très rarement à un évanouissement ou sensation d'évanouissement (It may rarely cause a fainting or a faint feeling).*

Explanation of abbreviations is done by adding their developed form. We exploit the lexicon with over 4,000 medical abbreviations and their developed forms. Hence, if we find an abbreviation in the text, we verify if the lexicon proposes its explanation. If so, the explanation is added in the simplified sentence between brackets:

- source: *L'OMS recommande un calendrier de vaccination antitétanique durant l'enfance de 5 doses. (WHO advises a five-injection calendar for the tetanus vaccine during childhood)*
- simplified: *L'OMS (Organisation mondiale de la santé) recommande un calendrier de vaccination antitétanique durant l'enfance de 5 doses. (WHO (World Health Organization) advises a five-injection calendar for the tetanus vaccine during childhood)*

2.2. Syntactic Simplification

Several situations trigger syntactic simplification. Long sentences containing several propositions are divided, which usually results in two simpler sentences. Three types of markers are used to determine where the sentence should be divided: coordination (example 1), relative (example 2), and discursive (example 3).

1. *L'administration concomitante du chlorhydrate de tamsulosine avec la paroxétine a entrainé une augmentation de la Cmax et de l'ASC du chlorhydrate de tamsulosine d'un facteur 1,3 et 1,6 respectivement, mais ces augmentations ne sont pas considérées comme étant cliniquement*

significatives (Concomitant administration of tamsulosin hydrochloride with paroxetine leads to the increase of Cmax and tamsulosin hydrochloride ASC of respectively 1.3 and 1.6, but this increase is not considered to be clinically significative). This sentence is split on *mais (but)* during the syntactic simplification.

2. *Le tramadol peut provoquer chez les nouveau-nés des modifications de la fréquence respiratoire, qui sont généralement sans conséquences cliniques préjudiciables (In newborn babies, tramadol can cause modifications of breath frequency, which do not have harmful clinical consequences).* This sentence is split on *qui (which)* and the subject is repeated anaphorically *(Elles (they))* in the second sentence.
3. *La ranitidine est éliminée par voie rénale, aussi les taux plasmatiques du médicament augmentent chez les patients présentant une insuffisance rénale (Ranitidine is eliminated by the kidney, also the plasma rate of the medication increases in patients suffering from renal insufficiency).* This sentence is split on *aussi (also)* and the second sentence starts with *De cette manière (In this way)* to keep the discursive relations intact.

We also transform passive sentences into active sentences. Hence, when we find a passive sentence, we extract the passive verbal phrase, the subject, and the object. The verb is then transformed into its active form, the subject becomes the object, and the object becomes the subject. This example illustrates this kind of transformations:

- source: *Une prudence particulire devra être observée par les conducteurs d'automobiles et les utilisateurs de machines (A particular attention should be paid by car drivers and vehicle users).*
- simplified: *Les conducteurs d'automobiles et les utilisateurs de machines devront observer une prudence particulire (Car drivers and vehicle users should be particularly attentive).*

2.3. Evaluation

The evaluation is done on 30 clinical cases (11,864 word occurrences and 755 sentences) from different specialties, such as provided by the CAS corpus [15]. The evaluation is done manually, since the automatic metrics of the simplification are not representative of the obtained quality and are often criticized [16]. Three evaluation criteria are considered:

- Simplicity: the evaluators are presented with two sentences (source and simplified) and are asked to indicate which sentence corresponds to the simplified version. Simplicity is evaluated as proportion of correctly found simplified sentences;
- Adequacy: two sentences (technical and simplified) are presented to the evaluators and the evaluators are asked to indicate the semantic similarity between these sentences on a Likert scale going from 1 (sentences are different semantically) to 5 (sentences are identical semantically);
- Grammaticality: in 100 simplified sentences, the evaluator verifies each type of transformations (gender, verb and determiner concordance, sentence splitting). The evaluator decides on the grammaticality of each transformation, which permits to compute the precision of the simplified sentence and represents the grammaticality obtained.

Grammaticality is evaluated by one person. Adequacy and Simplicity are evaluated by four people. None of the evaluators has expertise in medical domain.

3. Evaluation

Table 1. Evaluation of the grammaticality for different types of transformations.

Types of transformations	Nb. evaluated transformations	Nb. correct transformations	Precision
Gender	10	3	0.30
Number concordance	12	3	0.25
Verb concordance	3	2	0.75
Determiner concordance	64	50	0.78
Sentence division	10	9	0.90

Table 1 shows the results of manual evaluation of grammaticality according to different types of transformations performed. We can see that precision is high for verb and determiner concordance, and for the splitting of sentences. Yet, the gender and number of concordance of adjectives and past participles remains low. This may be due to errors in the syntactic parsing by Cordial, which we use for determining the gender and number of words.

Table 2. Results of simplicity and adequacy.

Type	Score
Simplicity	74%
Adequacy	4.05/5

Table 2 shows the results of the simplicity and adequacy evaluation. In 74% of the sentences, the evaluators determined correctly which sentence was the simplified version. We assume that the sentences wrongly indicated as simplified versions correspond to situations in which the simpler equivalents of technical terms are longer than the actual technical term. Therefore, the evaluators may have marked the technical version of the sentences because they are shorter. The evaluators gave the average score 4.05 (out of 5) for the adequacy, which is a rather high score. Simplified sentences judged as semantically different also contain simpler equivalents which are longer than the corresponding technical terms. In this case, long explanations or paraphrases may introduce extra information which decreases the semantic similarity between the sentences (source and simplified). We assume that our results show rather high evaluation scores. Yet, we can improve the grammaticality by defining better transformation and grammaticality rules. The simplicity score can be improved with a higher quality lexicon and with additional transformation rules. Finally, the adequacy scores can be improved if the quality of the lexicon is also improved. Our future work will address these issues.

4. Conclusion

We proposed a rule-based system to simplify technical texts in French from the medical domain. Our system performs both syntactic and lexical simplifications. The lexical simplification is done at two levels: adding the extended forms of abbreviations and substituting technical terms by their simpler equivalents. Syntactic simplifications are also performed at two levels: transforming passive sentences into active sentences and dividing long sentences. In addition, we perform fine-grained transformations for adjusting and checking the grammaticality of the simplified sentences. Our system was then evaluated according to three metrics: simplicity, adequacy and grammaticality. These metrics are evaluated manually. All metrics show rather high scores, which can be improved yet with a higher quality of the lexicon and of the transformation rules.

Acknowledgments. This work was partly funded by the French National Agency for Research (ANR) as part of the CLEAR project (*Communication, Literacy, Education, Accessibility, Readability*), ANR-17-CE19-0016-01. We also thank the evaluators for their help with the evaluation of the simplified sentences.

References

[1] Ruel J, Kassi B, Moreau AC, and Mbida-Mballa SL. *Guide de rdéaction pour une information accessible.* Pavillon du Parc, Gatineau, 2011.

[2] OCDE. *Guide de style de l'OCDE Troisième édition: Troisième édition*. OECD Publishing, 2015.

[3] UNAPEI. *L'information pour tous*. UNAPEI, 2019.

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

[5] Kim YS, Hullman J, Burgess M, and Adar E. SimpleScience: Lexical simplification of scientific terminology. In *EMNLP*, pages 1–6, 2016.

[6] Zhao S, Wang H, and Liu T. Leveraging multiple MT engines for paraphrase generation. In *COLING*, pages 1326–1334, 2010.

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

[8] Sennrich R, Haddow B, and Birch A. Improving neural machine translation models with monolingual data. In *Proc of the Ann Meeting of the Assoc for Comp Linguistics*, pages 86–96, Berlin, Germany, 2016.

[9] Nisioi S, Stajner S, Ponzetto SP, and Dinu LP. Exploring neural text simplification models. In *Ann Meeting of the Assoc for Comp Linguistics*, pages 85–91, 2017.

[10] Carroll J, Minnen G, Pearce D, Canning Y, Devlin S, and Tait J. Simplifying Text for Language-Impaired Readers. In *EACL*, pages 269–270, 1999.

[11] Bautista S, Gervás P, and Madrid RI. 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.

[12] De Belder J, Deschacht K, and Moens MF. Lexical simplification. In *ITEC*, 2010. 1-4.

[13] Laurent D, Nègre S, and Ségula P. L'analyseur syntaxique Cordial dans Passage. In *Traitement Automatique des Langues Naturelles (TALN),* 2009.

[14] Koptient A and Grabar N. Large rated lexicon for the simplification of medical texts. In *HEALTHINFO*, pages 1–7, 2020.

[15] Grabar N, Dalloux C, and Claveau V. CAS: corpus of clinical cases in French. *Journal of BioMedical Semantics*, 11(1):1–7, 2020.

[16] Sulem E, Abend O, and Rappoport A. BLEU is not suitable for the evaluation of text simplification. In *S*, pages 738–744, Brussels, Belgium, 2018.