Radiological Text Simplification Using a General Knowledge Base

Lionel Ramadier, Mathieu Lafourcade

Radiological text simplification using a general knowledge base

Lionel Ramadier[1] and *Mathieu Lafourcade*[2]

(1)Department of Radiology, University Hospital of Montpellier,
(2)LIRMM, University of Montpellier,
France
l-ramadier@chu-montpellier.fr
mathieu.lafourcade@lirmm.fr

Abstract. 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.

Keywords: NLP, BioNLP, text simplification

1 Introduction

Text simplification (TS) is a challenging natural language processing (NLP) task. It is an operation to simplify an existing corpus, texts or sentences while the underlying meaning and information remain the same. The main goal of TS is to make information more accessible to the large numbers of people with reduced literacy. TS can be viewed as an example of a monolingual translation task, where the source language needs to be translated into a simplified version of the same language.

Its application to the medical domain is of special importance. Understanding medical text might be particularly challenging for laymen readers who are not used to looking up unknown terms while reading. So, making record information available to the patients is a prioritized goal for many countries. It is crucial for patients to understand texts from the medical domain.

Medical texts are difficult to understand for non expert [1] because doctors often write with specialized terms (*ataxia*) and abbreviations (*HIV* for *human immunodeficiency virus*) which may require advance knowledge of medicine or

biology. There is a mismatch between the content delivered by medical practitioners and the consumers who have a limited health knowledge. Medical terms have been shown to be obstacles for patients [2], [3]. Moreover, medical reports are often written under time pressure by professionals for professionals. This results in a telegraphic style, with omissions, abbreviations,and sometimes misspellings [4].

However,there has been relatively limited prior research on tools to automate the simplification medical texts [5]. One tool that address this problem is a system built by Elhadad [6]. The author identifies difficult terms and retrieves definitions thanks to Google search engine. The tool improve reader's comprehension by an average 1.5 points on a 5 point scale. An other method [7] identified difficult terms in the document and try to simplify by replacing them with synonyms or by explaining them using easier words. The results reported correct simplification in 68 % of identified difficult word. For the French language, few studies deal with the issue of simplification of medical texts. A study [8] try to simplify a dialogue task between a virtual patient and a doctor.

The aim of our study is to simplify French radiology reports thanks to a general knowledge base that contains both general and specialized knowledge. For this issue, we use not only synonyms but also other hierarchically and/or semantically relates terms. In this paper, after a presentation of related work (section 2) we present our method (section 3) of semantic simplification thanks to a French lexical network (JeuxDeMots (JDM)), then we discuss experiments and analyze the results (sections 4 and 5).

2 Related work

The level of difficulty can vary between kind of medical texts [9], and even brochures for patients can be difficult to understand [10]. Medical texts, such as radiology reports, are characterized by sentences containing a lot of medical terms and a frequent use of abbreviation form. Previous studies [11] have shown that replacing difficult words with easier synonyms can reduce the level of difficult in a medical text. This synonym replacement method has been evaluated on medical English text [12],[15] and also on Swedish medical Text [11]. Semi-automatic adaption of word choice has been evaluated on English medical text [12] and automatic adaption on Swedish non-medical text [13]. Study used synonym lexicons and replaced difficult word with easier synonyms. The level of difficulty of a word was determined by measuring its frequency in a general corpus. In [7], the author used two sources of vocabulary knowledge : Unified Medical Language System [1] and the open-access collaborative (OAC) consumer health vocabulary (CHV). They employ two strategies to reduce the vocabulary difficulty of medical reports :

— synonym replacement

1. https://www.nlm.nih.gov/research/umls/

— explanation insertion

Leroy et al [12] developed an algorithm that uses term familiarity to identify difficult text and select easier alternatives from lexical resources such as WordNet, UMLS and Wiktionary. Their results show that term familiarity is a valuable component in simplifying text in an efficient manner.

For synonym replacement to be a meaningful method for text simplification, there have to exist synonyms that are near enough not to change the content of what is written. For describing medical concepts, there is, however, often one set of terms that are used by medical professionals, whereas another set of easier terms are used by patients [14].This means that synonym replacement could have a large potential for simplifying medical text. For English, there is a consumer health vocabulary initiative connecting laymen's expressions to technical terminology [5] as well as several medical terminology containing synonyms like MeSH [2] and SNOMED CT [3]. MeSH (Medical Subject Headings) is the National Library of Medicine's controlled vocabulary thesaurus, used for indexing articles for the MEDLINE database and SNOMED CT is one of a suite of designated standards for use in U.S. Federal Government systems for the electronic exchange of clinical health information.

In the radiology domain, several studies have shown that radiology reports are among the most difficult form of clinical text to understand [16]. The aim of a Swedish study [17] is to be able to develop a text simplification tool enabling patients to better understand text for a large corpus of Swedish radiology reports.

3 Our approach

We study simplification of one medical text genre,radiology reports. We use a replacement method by synonym but also by hierarchical relations. This latter are very useful because a term can be explained as a specific incidence of its parents. For example *hepatocellular carcinoma* is a *tumor of liver* or *pulmonary embolism* is a *lung disease*. The knowledge base on which our radiology reports simplification relies is the French lexical network JDM [4] [18].

3.1 Resources

The JeuxDeMots Lexical Network

JDM network is a lexical-semantic graph for the French language whose lexical relations are generated both through GWAP (Games With A Purpose) and via a contributory tool called Diko (manual insertion and automatic inferences

2. www.nlm.nih.gov/mesh/
3. http://www.snomed.org/snomed-ct
4. http://www.jeuxdemots.org/

with validations) [19]. At the time of this writing (February, 2017), the JDM network contains over 67 millions of relations between around one million of terms. The following table provides an order of size of the amount of information we have at our disposal about the radiology areas (table 1).

Terms	Outgoing links	Incoming links
medicine	22 108	24 100
anatomy	10 477	11 453
radiology	382	502
medical imaging	541	556

Table 1. Number of relations of some key terms within the JDM lexical network.

It exists 80 lexico-semantic relations into the network but in this work we use only three different relations in addition to lexical information.

r_synonym synonyms or quasi-synonyms
r_syn_strict strict synonyms
r_isa generic term
r_equiv acronym or abbreviation

Fig. 1. the relations used for medical simplification text

We use this network because it combines weight and annotations [20] on typed relations between terms. In the network JDM, the relations are weighted, the weight reflects the strength of association between terms. In specialized knowledge, the correlation between the weight of the relation and its importance is not strict.This is why it appears interesting to use annotations for some relations as they can be of a great help in the medical area. This annotations could help us in the task of text simplification (thanks to the annotation *ordinary language*) (figure 3). A given relation to be annotated is reified (represented by a specific node) and this node is associated to various annotations and any other regular terms. The annotation relation type is a kind of relation among others figure 2.

The corpus of radiology reports

The corpus contains 30 000 radiology reports, from different institutions, concerning the different medical imaging techniques (MRI, scanner, medical ultrasound, X-ray radiology, vascular radiology, etc). These reports are written in semi-structured way. They are generally divided into four parts. Each part is

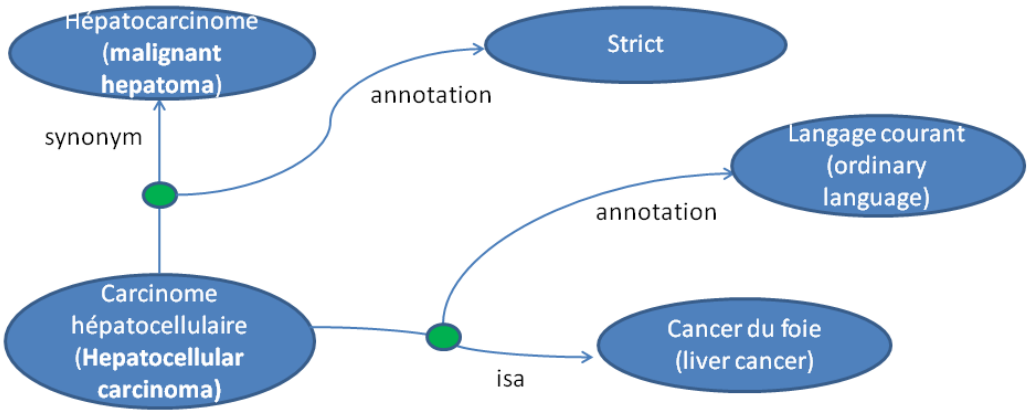

Fig. 2. A given relation to be annotated is reified (represented by a specific node, here with green circles) and this node is associated to various annotations and any other regular terms.

written by the radiologist in a very free style, often with a profusion of acronyms and specialized terms. The records are deidentified with anonymized serial numbers for individual patients. The reports are examinations of all patients for this period i.e both genders and all ages from babies to a 98 year-old.

3.2 Method

To support patient radiology report comprehension, it is important to identify words that matter most to patients in their reports. The identification of the compound terms is made compared to the content of JDM, in a first step. We use the underscore to separate the two parts of a compound word so that it is considered as an entity at the time of the extraction (tibia_fracture).

In a second step, we extract difficult term by using traditional methods i.e term and document frequency (TF and DF) to calculate the IDF (Inverse Document Frequency). For each difficult term, we look at to the content of JDM

Fig. 3. Example of term *prurit* (*pruritus*) with annotations between brackets. Several annotations are possible for a given relation like *ordinary language* (*langage courant* in French)

for the synonym or hyperonym relation. A difficult word can have several synonyms and sometimes this latter are not easier. For instance, in French language the term *carcinome hepatocellulaire* (*Hepatocellular carcinoma*) has for synonym *hepatocarcinome*. In this case, we can look at the hyperonym relation. To choose the right easier term, we use relation annotations [21]. If the relation has an annotation (*langage courant* in French or *ordinary language*), then the original words are replaced by pertinent synonym or hyperonym. For instance, the term *aphasia* will be replaced by *mutism* because in the network JDM *aphasia* r_synonym *mutism* (ordinary language) (Figure 4). The multi-word *hepatocellular carcinoma* will be replaced by the hyperonym *liver cancer* thanks to the network JDM (hepatocellular carcinoma r_isa *liver cancer* (*ordinary language*)). In this case, we have chosen the hyperonym relation because the different synonym of hepatocellular carcinoma are not easier to understand. We systematically replace the term *anterior* by *front of* and *posterior* by *behind*. Some abbreviations (e.g "MVC") will be replaced with their more understandable full names (e.g "motor vehicle collision"), but the abbreviations AIDS will not be replaced by their full names because the patient know and understand this abbreviation. We can distinguish the two case, because in the network the relation annotation allows to distinguish both cases.

If the difficult term has not a relation annotation, we apply another approach. If the synonym or hyperonym are compound terms like *tumeur du cerveau* (*brain*

le patient a une *aphasie* depuis deux jours. (The patient has an *aphasia* for two days) Le patient a un **mutisme** depuis deux jours(The patient has a **mutism** for two days)
Le patient se plaint de *céphalée* (The patient complains of cephalgia). Le patient se plaint de **maux de tête** (The patient complains of *headache*).
Symptôme : *hématurie* (symptom:*hematuria*) Symptôme : **sang dans les urines** (symptom:**blood in the urine**)

Fig. 4. Example of sentence translation. The replaced terms are on bold

tumor) do not have relation annotations, we extract semantic information for each word that makes up the compound terms from the JDM network. Indeed, lexical information indicates whether a word is part of the common language or not. For instance, glioblastoma (figure 5) is a brain tumor (hyperonym relation). As this hyperonym relation has not an annotation, we extract for *brain* and *tumor*, semantic information from JDM. Each word belongs to the ordinary language, then we replace *glioblastoma* by *brain tumor*.

Fig. 5. Example of term *glioblastome* (*glioblastoma*) with annotations between brackets. Several annotations are possible for a given relation.

4 Experiment and Results

4.1 Experiment

We use a corpus subset (200 radiology) reports, and we simplified them using our method (thanks to the network JDM).

For the manual evaluation, 250 sentences were randomly selected for human review and cloze testing (a standard comprehension test procedure) [22]. An expert reviewed the translations for corrections. According to the standard cloze procedure, every 5th word of each report was replaced with a blank space.

We have recruited 4 persons, who were not doctors but highly educated (1 at undergraduate school level and 3 at graduate school level), to evaluate the system. Each subject has evaluated the original and simplified reports. They were asked to fill in the blank spaces.

We calculate for each report a cloze score which is the percentage of answers that matched with the deleted word. We compared the average cloze scores of the original and translated radiology reports.

4.2 Results

On average, 10.6 terms were simplified in reports. Most of the simplifications (75%) were deemed correct by an expert reviewer. For 12% of the sentences, the replaced word has a slightly different meaning for the original word. This errors can explain because sometimes the synonym was not strict. For instance a cyst (kyste in French) and abscess (abcès in French) are synonym or quasi-synonym in the network but in the field on medicine, the meaning are different. We show some words typical for a professional language that have been replaced with every day French words, or abbreviations that have been replaced by an expanded form (Table 2).

Original terms	**Replaced with**
aphasie (aphasia)	mutisme (mutism)
céphalée (cephalgia)	maux de tête (headache)
prurit (pruritus)	démangeaison (itch)
dyspnée (dyspnea)	difficulté à respirer (shortness of breath)
glioblastome (glioblastoma)	tumeur maligne du cerveau (brain tumor)
CHC (hepatocellular carcinoma)	cancer du foie (liver cancer)
arthrite (arthritis)	inflammation des articulations (joint-inflammation)
TS (SA)	tentative de suicide (suicide attempt)

Table 2. Examples of replaced terms.

If the cloze score is between 50-60%,then the document should be readable. In table 3 and table 4 , we show the results for original and simplified reports.

Original reports	**Simplified reports**
18%	48%

Table 3. Cloze score for original and simplified reports using only annotation relation .

Original reports	**Simplified reports**
18%	57%

Table 4. Cloze score for original and simplified reports using annotation and semantic information.

The cloze score of the original radiology reports (18%) indicate that these documents are difficult for lay people to understand.

5 Discussion

We describe a text simplification system for a French radiology corpus. The need to improve the understanding of medical reports for patient is important. The patient want more and more to understand the different medical records. The radiology reports are the most difficult to understand for the consumers.The cloze score for the original report is lower than other study [22] that deal with various medical reports(discharge summaries, surgery report, only one radiology reports). We have implemented a prototype to improve the readability for lay readers. This study focused on vocabulary difficulty. Our method relies on the JDM network to try to simplify difficult words. Indeed, to choose an easier term, we make use of relation annotations present in the lexical semantic network. This method allows us to choose the right term easily. 80% of replaced term seem helpful with the same meaning. If we use only the relation annotations for the task of simplification, we get a cloze score of 48%. If we use the second approach based on semantic information we improve our results and we reach a cloze score of 57%.

But 35% of difficult terms are not replaced because they have not annotations (ordinary language) in the network. It needs to improve the coverage of the annotations inside the network to reach better results. The manual evaluation also showed that the original semantic meaning had been slightly altered in some sentences. In some case, some words are not strictly synonym and the

replacement involve a slightly change of meaning. For instance, the replacement oedema by swelling entail a change of meaning. Moreover,in order to include abbreviations and acronyms in the synonym replacement method studied here, an abbreviation disambiguation needs to be carried out first. An acronym or an abbreviations can have two different meanings in the field of medicine.

The average cloze test of the simplified reports are high, it reach score 50-60% to be fairly readable. Our results are close to those [22] although our corpus is larger and contains only radiology reports.

Our system needs much improvement. We intend to simplify the syntax. Another task is to improve the coverage of the annotation relations (*ordinary language*) inside the network.

6 Conclusion

We have developed a system which goal is to improve patient comprehension. The results presented here are preliminary but are very promising. In this work, we have used the JeuxDeMots lexical-semantic network as a support of knowledge. Although this network is general, it contains many specialty data, including medicine/radiology that may helpful in the simplification task framework.

The difficulty of a word was assessed by the presence or not of relation annotations in the network JDM. It seems a good way to evaluate the difficulty of a word. The replacement was mainly evaluated by the cloze test. Studies on a larger reader group are required to draw any conclusions on the effect of our method for assessment of simplification. We have to recognize errors in order to eliminate them. An another future improvement is to use the definitions present in the network in order to generate explanations.

In a future work, another challenge is to simplify the syntax of radiology reports. A previous study [23] showed significant differences in syntactic content and complexity between medical discharge summaries and everyday English papers. An other survey emphasized the difficulty of syntactic text simplification [24]. For this task, we would be able to realize a grammar simplification (for instance, long sentences were broke down into two or more shorter sentences).

We also plan to test our approach in other medical domains ,such as for example the oncology, because JDM contains data about this domain.

References

1. Keselman, A., Smith, C. A. (2012). A classification of errors in lay comprehension of medical documents. Journal of biomedical informatics, 45(6), 1151-1163.
2. Chapman, K., Abraham, C., Jenkins, V., and Fallowfield, L. (2003). Lay understanding of terms used in cancer consultations. Psycho-Oncology, 12(6), 557-566.

3. Lerner, E. B., Jehle, D. V., Janicke, D. M., and Moscati, R. M. (2000). Medical communication: do our patients understand?. The American journal of emergency medicine, 18(7), 764-766.
4. Hagege, C., Marchal, P., Gicquel, Q., Darmoni, S., Pereira, S., Metzger, M. H. (2010, August). Linguistic and temporal processing for discovering hospital acquired infection from patient records. In International Workshop on Knowledge Representation for Health Care (pp. 70-84). Springer Berlin Heidelberg.
5. 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.
6. Elhadad, N. (2006). Comprehending technical texts: predicting and defining unfamiliar terms. In AMIA.
7. Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A., Rosendale, D. (2007, November). Making texts in electronic health records comprehensible to consumers: a prototype translator. In AMIA (pp. 846-50).
8. Pierre, L. C. L. D. B., Rosset, Z. S.(2016) Managing Linguistic and Terminological Variation in a Medical Dialogue System. LREC Portoroz (pp. 3167-3173).
9. 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.
10. Kokkinakis, D., Forsberg, M., Kokkinakis, S. J., Smith, F., Öhlen, J. (2012, September). Literacy Demands and Information to Cancer Patients. In International Conference on Text, Speech and Dialogue (pp. 64-71). Springer Berlin Heidelberg.
11. Skeppstedt, E. A. T. F. M., Kvist, M. (2014, April). 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)@ EACL (pp. 57-65).
12. Leroy, G., Endicott, J. E., Mouradi, O., Kauchak, D., Just, M. (2012). Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In AMIA.
13. Keskisarkka, R. (2012). Automatic text simplification via synonym replacement.In Proceedings of Swedish Language Technology Conference.
14. Kokkinakis, D., and Gronostaj, M. T. (2006). Lay language versus professional language within the cardiovascular subdomain–a contrastive study. Proc. of BIO'06.
15. 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.
16. Keselman, A., Slaughter, L., Arnott-Smith, C., Kim, H., Divita, G., Browne, A.,Zeng-Treitler, Q. (2007). Towards consumer-friendly PHRs: patients' experience with reviewing their health records. In AMIA Annual Symposium Proceedings (Vol. 2007, p. 399). American Medical Informatics Association.
17. 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 (pp. 55-59). Linköping University Electronic Press.

18. Lafourcade, M. (2007, December). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In SNLP'07: 7th international symposium on natural language processing (p. 7).

19. Lafourcade, M., Joubert, A., , and Le Brun, N. (2015). Games with a Purpose (GWAPS).John Wiley and Sons.ISBN: 978-1-84821-803-1.

20. Ramadier, L., Zarrouk, M., Lafourcade, M., Micheau, A. (2014, April). Spreading relation annotations in a lexical semantic network applied to radiology. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 40-51). Springer Berlin Heidelberg.

21. 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.

22. Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A., Rosendale, D. (2007, November). Making texts in electronic health records comprehensible to consumers: a prototype translator. In AMIA (pp. 846-50).

23. Campbell, D. A., and Johnson, S. B. (2001). Comparing syntactic complexity in medical and non-medical corpora. In Proceedings of the AMIA Symposium (p. 90). American Medical Informatics Association.

24. Kandula, S., Curtis, D., and Zeng-Treitler, Q. (2010, November). A semantic and syntactic text simplification tool for health content. In AMIA Annu Symp Proc (Vol. 2010, pp. 366-70).