French Biomedical Text Simplification: When Small and Precise Helps

Rémi Cardon, Natalia Grabar
UMR CNRS 8163 – STL, Université de Lille
F-59000 Lille, France
remi.cardon@univ-lille.fr, natalia.grabar@univ-lille.fr

Abstract

We present experiments on biomedical text simplification in French. We use two kinds of corpora – parallel sentences extracted from existing health comparable corpora in French and WikiLarge corpus translated from English to French – and a lexicon that associates medical terms with paraphrases. Then, we train neural models on these parallel corpora using different ratios of general and specialized sentences. We evaluate the results with BLEU, SARI and Kandel scores. The results point out that little specialized data helps significantly the simplification.

1 Introduction

The goal of text simplification is to make complex texts accessible to a given target audience (children, foreigners, diseased people...) or to make complex texts to be more easily processed by NLP applications. There is little existing work on biomedical text simplification in English (Peng et al., 2012; Shardlow and Nawaz, 2019), and on text simplification for the general language in French (Abdul Rauf et al., 2020; Gala et al., 2020; Sauvan et al., 2020; Brouwers et al., 2014).

Research on text simplification is usually performed on open domain English. Currently used methods rely on deep learning approaches and require large parallel monolingual corpora in which one complex sentence is paired with one or more simplified versions (Nisioi et al., 2017; Cooper and Shardlow, 2020). Two English datasets are commonly used: (1) Newsela (Xu et al., 2015), a corpus with news articles that are manually re-written according to four levels of simplification, and (2) WikiLarge (Zhang and Lapata, 2017), issued from a compilation of three previously released simplification corpora all extracted from Wikipedia (Zhu et al., 2010; Woodsend and Lapata, 2011; Kauchak, 2013). The availability of corpora and resources plays indeed a very important role and condition the feasibility of the NLP research.

We build and test models for biomedical text simplification in French. We first describe the data used (Section 2) and various configurations of the experiments (Section 3). We then indicate the evaluation principles (Section 4), and present and discuss the results (Section 5). We also provide the WikiLarge FR corpus and a set of native parallel sentences in French, from general and biomedical languages.

2 Linguistic Data

	Total		*Train*		*Validation*		*Test*	
Corpus	*Pairs*	*Tokens*	*Pairs*	*Tokens*	*Pairs*	*Tokens*	*Pairs*	*Tokens*
WikiLarge FR	297,494	12,753,567	296,402	12,695,192	992	42,676	100	4,302
CLEAR	4,596	226,149	4,196	206,500	300	7,381	100	4,965

Table 1: Size of the two parallel corpora

We use two sets of parallel corpora dedicated to the simplification. One set is obtained from the freely available simplification corpus CLEAR in French (Grabar and Cardon, 2018). This is a comparable

Proceedings of the 28th International Conference on Computational Linguistics, pages 710–716
Barcelona, Spain (Online), December 8-13, 2020

corpus, which contains three types of texts (medical literature reviews, drug information, and medical articles from Wikipedia and Vikidia), and from which parallel sentences were extracted. This parallel corpus contains 4,596 sentence pairs, which is not sufficient when using the current state-of-the-art methods, that rely on deep learning. Another corpus is obtained thanks to the automatic translation of WikiLarge in French. The translation has been done using OpenNMT-py (Klein et al., 2017) with the default parameters, and the En-Fr model provided. This WikiLarge FR parallel corpus contains almost 300,000 sentence pairs. Table 1 indicates the volume of data in both corpora. We segmented the CLEAR corpus into train, validation and test sets: 100 examples for testing, three times as many for validation, and the rest for training. WikiLarge FR is already segmented in these three sets, yet we decided to reduce the WikiLarge FR test set from 359 pairs to 100 to make it comparable with the CLEAR test set. As these two corpora contain data from Wikipedia, we checked for duplicates to avoid having identical pairs in two different sets, and we found none.

We also use a lexicon that proposes layman paraphrases for technical medical terms, like {*hypotension*; *baisse de la tension artérielle*} ({*hypotension*; *decrease in arterial pressure*}). This lexicon has been built using medical terminologies (Lindberg et al., 1993) and corpora (CLEAR corpus and various discussion fora) in French. The lexicon currently contains 7,580 paraphrases for 4,516 medical terms.

3 Experimental Protocol

The purpose of the experimental protocol is to focus on two aspects:

1. *Impact of the general and medical language corpora*, for which we use different ratios of WikiLarge FR and CLEAR sets. Since we aim at the simplification of biomedical texts and since the CLEAR corpus is not large, we always use all of the CLEAR training and validation sets. Then, we gradually add WikiLarge FR examples to the training set: we start with a model trained on the same number of examples from both corpora (ratio 1:1), and increase examples from WikiLarge FR respecting the ratios 1:5, 1:10, 1:25, 1:50, and up to 1:75. The 1:75 ratio value is an approximation, it corresponds to the case where the entire WikiLarge FR training set is used in the experiment. Each additional set with WikiLarge FR examples is selected randomly and added to the previous set. These experiments permit to observe the robustness of the models when tested on biomedical and general language, and to observe whether these additional examples improve the results;

2. *Impact of lexicon*. We perform the same experiments to which we feed the lexicon. The lexicon is exploited in two ways: (1) during the simplification phase with the OpenNMT-py `--phrase_table` flag that it usually used for dealing with unknown words. That flag was used in the same way in a work for simplification of clinical letters in English (Shardlow and Nawaz, 2019) and (2) during the training phase for which we add the entire lexicon to the training set.

We refer to the first set of experiments as NPT (no phrase table), the second set of experiments as PTS (phrase table used in the simplification phase) and the third set of experiments as PTT (phrase table added to the training set). Across each series of experiments, the validation and test sets are the same, while the training sets differ in the ratios and whether or how the lexicon is used.

We use OpenNMT-py for the sentence simplification with the following configuration: two bidirectional LSTM layers of 500 units for the encoder and the decoder, ADAM optimizer, learning rate 0.001, dropout probability 0.3, attention dropout probability 0.2. Each individual training took about five hours on a Geforce RTX 2070 GPU. During the simplification phase, we use the `--replace_unk` flag which tells the program to copy unknown words from the input to the output, except for the PTS set of experiments where it is replaced by the `--phrase_table` flag that uses the lexicon.

We report three baselines: (1) one model trained on CLEAR only, (2) another model trained on WikiLarge FR only, and (3) the identity baseline which corresponds to the case where the output is the copy of the input.

4 Evaluation

We evaluate the models using several metrics: (1) BLEU (Papineni et al., 2002), initially designed for the evaluation in machine translation, is also used in text simplification which can be seen as a monolingual translation task. It compares the system output with the reference data. This metric gives a rough indication of the performance of a system, especially regarding grammaticality and meaning preservation, but it is not a strong indicator for simplification (Sulem et al., 2018); (2) SARI (Xu et al., 2016) is currently considered as the most common metric for text simplification. SARI is computed by comparing the system output against the reference, and against the input as well. It should be noted that SARI is more reliable when several references are available (Alva-Manchego et al., 2020; Zhang and Lapata, 2017), which is not the case in our experiments; (3) Kandel (Kandel and Moles, 1958) is a readability metric. It does not compare the output with the reference or the input, and only provides formal indicators such as sentence length and number of syllables per word. It is an adaptation of the Flesch (Flesch, 1948) readability measure – that was designed for English – to the French language. The absolute indexes are not informative by themselves: the measure is described to be more relevant for comparisons. Higher scores mean that the text should be easier to read.

We computed the first two metrics with the EASSE evaluation suite for automatic text simplification (Alva-Manchego et al., 2019).

5 Quantitative and Qualitative Results

	WikiLarge FR			*CLEAR*		
Models	*BLEU*	*SARI*	*Kandel*	*BLEU*	*SARI*	*Kandel*
Identity baseline	60.02	25.05	81.15	55.00	23.73	76.67
WikiLarge FR	39.08	37.61	89.71	9.72	30.97	95.58
CLEAR	0.15	20.52	94.32	21.59	22.07	84.15
NPT 1:1	5.83	25.60	**98.20**	26.23	38.10	**84.26**
NPT 1:5	14.82	30.38	96.23	29.86	39.20	80.43
NPT 1:10	33.74	35.01	92.97	41.05	38.32	80.02
NPT 1:25	25.88	34.44	92.26	37.24	**40.34**	78.12
NPT 1:50	44.48	**38.93**	90.52	49.16	35.36	79.09
NPT 1:75	**49.67**	38.02	89.71	**50.23**	33.91	79.11
PTS 1:5	15.06	30.28	**103.29**	30.00	39.10	82.06
PTS 1:10	33.70	35.12	102.17	40.29	38.32	79.42
PTS 1:25	26.16	34.44	99.84	37.17	**40.09**	79.04
PTS 1:50	44.49	**39.05**	100.63	**48.16**	35.33	**89.08**
PTS 1:75	**49.70**	38.26	97.76	47.61	34.27	78.43
PTT 1:5	23.98	33.68	95.56	39.07	**40.94**	87.36
PTT 1:10	30.94	34.05	94.61	38.17	36.38	86.72
PTT 1:25	**37.29**	34.74	91.40	42.92	39.14	88.22
PTT 1:50	32.68	**36.73**	**98.81**	**49.72**	37.52	90.60
PTT 1:75	34.20	36.47	89.05	40.16	38.58	**92.35**

Table 2: Evaluation metrics for the various experiments obtained on WikiLarge FR and CLEAR

Table 2 indicates the SARI, BLEU and Kandel scores when testing the models on WikiLarge FR and CLEAR test sets. The first three rows show the baseline results. According to the Kandel scores, medical sentences from CLEAR are indeed less readable than sentences from Wikipedia. We can also see that training on CLEAR performs very poorly on the general language (<1 BLEU and 20.52 SARI). The performances are also poor on medical language but BLEU is way higher than on the general language (21.59 vs 0.15). This is due to the fact that the model trained on WikiLarge FR only performs quite well on WikiLarge FR (39.08 BLEU score) but poorly on CLEAR (9.72 BLEU score), which indicates

Models	*Test on WikiLarge FR*
Source	Le 14 octobre 1960, le candidat à la présidence John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan. (On October 14, 1960, presidential candidate John F. Kennedy proposed the concept of what became the Peace Corps on the Union Steps of Michigan.)
Reference	John F. Kennedy, un candidat à la présidence, a proposé l'idée de ce qui devint le Peace Corps sur les marches de l'Union du Michigan le 14 octobre 1960. (John F. Kennedy, a presidential candidate, proposed the idea of what became the Peace Corps on the Union steps of Michigan on October 14, 1960.)
Wikilarge	*No changes made*
CLEAR	le cancer est de la médecine (cancer is medicine)
NPT 1:25	En 1960, le candidat du président John F. Kennedy a suggéré le Peace Corps sur les marches de l'Union du Michigan. (In 1960, President John F. Kennedy's candidate suggested the Peace Corps on the Union steps of Michigan.)
NPT 1:50	En 1960, le candidat à la présidence John F. Kennedy a proposé l'idée du Peace Corps sur les marches de l'Union du Michigan. (In 1960, presidential candidate John F. Kennedy proposed the idea of the Peace Corps on the Union Steps of Michigan.)
PTT 1:75	John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan. (John F. Kennedy proposed the concept of what became the Peace Corps on the Union steps of Michigan.)

Table 3: Examples for simplification of sentences from WikiLarge FR

Models	*Test on CLEAR*
Source	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 4.2) (arterial hypotension may be observed if intravenous administration is too rapid, less than 60 minutes (see section 4.2).)
Reference	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide , inférieure à 60 min (arterial hypotension may be observed if intravenous administration is too rapid, less than 60 min.)
WikiLarge	Une artérielle artérielle peut être observée en cas de crise intraveineuse trop rapide et inférieure à 60 minutes ((An arterial arterial may be observed if intravenous crisis is too rapid and less than 60 minutes)
CLEAR	le traitement de la naissance de la naissance de l' repos [...] de l' repos de la médecine [...] de la médecine de la peau [...] de la peau de la (treatment of the birth of the birth of the rest [...] of the rest of the medicine [...] of the medicine of the skin [...] of the skin of the)
NPT 1:50 & 1:75	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (arterial hypotension may be observed if intravenous administration is too rapid, less than 60 minutes)
PTS 1:75	une tension inférieure à la normale artérielle peut être observée en cas d' administration intraveineuse trop rapide , inférieure à 60 minutes (lower than normal blood pressure may be observed if intravenous administration is too rapid, less than 60 minutes.)
PTT 1:50	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique « 3) (arterial hypotension may be observed if intravenous administration is too rapid, less than 60 minutes (see section " 3))
PTT 1:75	une diminution de la tension artérielle peut être observée en cas d' administration intraveineuse trop rapide , inférieure à 60 minutes (a decrease in blood pressure may be observed if intravenous administration is too rapid, less than 60 minutes)

Table 4: Examples for simplification of sentences from CLEAR

that the model trained on general language cannot handle domain specific data. When data from two corpora are used for the training, BLEU and SARI improve to up to 49.7 BLEU (*PTT 1:75*) and 39.05 SARI (*PTS 1:50*) on WikiLarge FR. This indicates that the volume of data is crucial. The input lexicon has little effect when used during the simplification phase. Yet, with several models, the lexicon proves to be efficient when used during the training phase. When testing on CLEAR, BLEU and SARI scores are higher: up to 50.23 BLEU (*NPT 1:75*) and 40.94 SARI (*PTT 1:5*). Here, all models perform better than the baselines: with as few as ∼4,000 examples of domain-specific data (for *NPT*) and ∼10,000 examples including the lexicon (*PTS* and *PTT*) this is a substantial improvement. Finally, as the Kandel index advantages models that output short sentences with no consideration to their contents, we will simply observe that no model worsens the Kandel readability score of the original texts – no model produces an output with a Kandel index that is lower than the identity baseline.

Table 3 shows some simplification examples provided for WikiLarge FR, and Table 4 shows some simplification examples for CLEAR.

In Table 3, the sentence *Le 14 octobre 1960, le candidat à la présidence John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan. (On October 14, 1960, presidential candidate John F. Kennedy proposed the concept of what became the Peace Corps on the Union Steps of Michigan.)* is processed. the baseline models provide either no changes (WikiLarge FR) or changes that are not meaningful (CLEAR). Indeed, the CLEAR baseline model applied to the WikiLarge FR test example proposes a grammatical sentence, but which has no semantic relation to the input (*cancer is medicine)*. This draws attention to the fact that simplification is very sensitive to the training data and needs caution. We can see that the CLEAR baseline model only outputs words related to the medical domain, regardless of the input. Hence, real improvement in quality is obtained with other models. Indeed, except the CLEAR baseline, other examples for WikiLarge correspond to the state-of-the-art transformations.

As for CLEAR (Table 4), we illustrate the modifications on a sentence that comes from drug information released by the French Ministry of Health: *une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 4.2). (arterial hypotension may be observed if intravenous administration is too rapid, less than 60 minutes (see section 4.2).)*. The source document is aimed at physicians, whereas the reference is aimed at patients and can be found in drug boxes. The only changes in the reference are the truncation of "minutes" and the deletion of the mention of another section. *NPT* examples are close to the reference except for the truncation, while *PTS* and *PTT* are more creative: *PTS 1:75* transforms *hypotension* in *lower than normal blood pressure*, while *PTT 1:75* transforms *hypotension* in *decrease in blood pressure*. Lexically, these are correct transformations and an improvement over the reference. Yet, by mechanically replacing *hypotension*, *PTS* creates an ungrammatical sentence in French.

6 Conclusion

We addressed the biomedical text simplification in French, which is, to our knowledge, the first attempt to perform this task with a machine translation technique. In order to cope with the lack of French data for simplification, we translated to French a freely available English corpus. We made the translation available, as well as the parallel data from CLEAR[1]. Using those data we achieved improvements over the baselines for French biomedical text simplification. Indeed, baselines produce incorrect and imperfect simplifications, but the results are significantly improved with large datasets.

We propose several experiments that indicate that (1) an automatically translated resource can help in a low resource setting, (2) a small amount of good quality specialized data can significantly improve overall performances, (3) multiword units processing can be dealt with by adding a lexicon to the training set. That is to say that mixing data of different linguistic nature helps simplification. Overall, we can obtain interesting simplification results which often prove to be more creative than the reference and yet correct. Besides, the impact of large native data for simplification should also be studied.

[1] `http://natalia.grabar.free.fr/resources.php`

Acknowledgements

This work was 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. The authors would like to thank the reviewers for their helpful comments and questions that permitted to improve the overall quality of the paper.

References

Sadaf Abdul Rauf, Anne-Laure Ligozat, François Yvon, Gabriel Illouz, and Thierry Hamon. 2020. Simplification automatique de texte dans un contexte de faibles ressources. In Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider, editors, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 332–341, Nancy, France, June. ATALA.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China, November. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. Syntactic sentence simplification for french. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 47–56.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5588–5594.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):p221 – 233, June.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Language Resources and Evaluation for Language Technologies (LREC)*, Marseille, France, May.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands, November. Association for Computational Linguistics.

L Kandel and A Moles. 1958. Application de l'indice de flesch à la langue française. *The Journal of Educational Research*, 21:283–287.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria, August. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

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

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Y. Peng, C. O. Tudor, M. Torii, C. H. Wu, and K. Vijay-Shanker. 2012. iSimp: A sentence simplification system for biomedical text. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–6.

Lauren Sauvan, Natacha Stolowy, Carlos Aguilar, Thomas François, Núria Gala, Frédéric Matonti, Eric Castet, and Aurélie Calabrèse. 2020. Text simplification to help individuals with low vision read more fluently. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 27–32, Marseille, France, May. European Language Resources Association.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389, Florence, Italy, July. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium, October-November. Association for Computational Linguistics.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China, August. Coling 2010 Organizing Committee.