Simplification automatique de textes biomédicaux en français : les données précises de petite taille aident

Rémi Cardon[1] Natalia Grabar[1]
(1) CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
{remi.cardon, natalia.grabar}@univ-lille.fr

RÉSUMÉ

Nous présentons un résumé en français et un résumé en anglais de l'article (Cardon & Grabar, 2020), publié dans les actes de la conférence *28th International Conference on Computational Linguistics (COLING 2020)*.

ABSTRACT

French Biomedical Text Simplification : When Small and Precise Helps

We present a French abstract and an English abstract of the article (Cardon & Grabar, 2020), published in the proceedings of the 28th International Conference on Computational Linguistics (COLING 2020).

MOTS-CLÉS : simplification automatique de textes, domaine biomédical.

KEYWORDS: automatic text simplification, biomedical domain.

1 Résumé en français

Nous présentons des expériences de simplification automatique de textes biomédicaux en français. Nous travaillons au niveau de la phrase. Dans ce travail, nous utilisons deux corpus :

1. 4 596 couples de phrases parallèles extraites automatiquement à partir de corpus comparables du domaine de la santé en français (Cardon & Grabar, 2019),
2. 297 494 couples de phrases parallèles issues du corpus de langue générale WikiLarge (Zhang & Lapata, 2017), dédié à la simplification, que nous avons traduit automatiquement de l'anglais vers le français.

Pour effectuer la simplification automatiquement, nous utilisons l'outil OpenNMT-py (Klein *et al.*, 2017), créé à l'origine pour la traduction bilingue. Le fonctionnement de cet outil est basé sur une architecture encodeur-décoder avec mécanisme d'attention. Nous exploitons OpenNMT-py pour transformer un texte technique en un texte simplifié. Nous entraînons des modèles neuronaux sur les corpus parallèles constitués, en utilisant différents ratios de phrases de langue générale et spécialisée. En effet, nous avons un volume de phrases assez élevé pour décrire la simplification de la langue générale. Cependant, ces phrases ne décrivent pas bien les transformations requises pour simplifier la langue médicale. Les phrases parallèles provenant du domaine biomédical permettent donc de combler cette limite. Nous utilisons aussi un lexique qui apparie des termes médicaux complexes avec des paraphrases accessibles au grand public (7 580 paraphrases pour 4 516 termes médicaux). Nous pouvons ainsi mener trois séries d'expériences :

1. le lexique de paraphrases n'est pas utilisé et la simplification est uniquement basée sur les exemples provenant des corpus d'entraînement ;
2. le lexique est exploité lors de la phase de simplification, où il sert à indiquer au modèle comment remplacer les termes inconnus, qui se trouvent dans le lexique de paraphrases ;
3. le lexique est exploité lors de la phase d'entraînement, où il est ajouté à l'ensemble d'entraînement, ce qui permet de compléter les données des corpus.

Nous évaluons les résultats avec les métriques BLEU (Papineni *et al.*, 2002), SARI (Xu *et al.*, 2016) et Kandel (Kandel & Moles, 1958). Globalement, les résultats indiquent que des données spécialisées, même en petite quantité, aident significativement la simplification.

2 English Abstract

We present experiments on automatic biomedical text simplification in French. We work at the sentence level. In this work, we use two corpora :

1. 4 596 parallel sentence pairs automatically extracted from a French biomedical corpus (Cardon & Grabar, 2019),
2. 297 494 parallel sentence pairs obtained from general language corpus WikiLarge (Zhang & Lapata, 2017), which we have automatically translated from English to French.

In order to perform automatic simplification, we use the OpenNMT-py tool (Klein *et al.*, 2017). It was created for machine translation. This tool operates on an encoder-decoder architecture with an attention mechanism. We exploit OpenNMT-py to transform technical sentences into simpler sentences. We train neural models on the parallel corpora, using different ratios of general language and specialized language. Indeed, the volume of data is sufficient for describing general language simplification. Though, the sentences do not describe transformations that are specific to the medical domain. The parallel sentences from the medical domain allow us to fill this gap. We also use a lexicon that maps complex medical terms with laymen paraphrases (7 580 paraphrases for 4 516 medical terms). Thus we can perform three series of experiments

1. the lexicon is not used and the simplification is only based on the examples from the training corpora ;
2. the lexicon is exploited during the simplification phase, where it is used to indicate to the model how to substitute unknown terms that are present in the lexicon ;
3. the lexicon is exploited during the training phase, where it is added to the training set, where it complements the parallel corpora.

We evaluate the results with three metrics : BLEU (Papineni *et al.*, 2002), SARI (Xu *et al.*, 2016) and Kandel (Kandel & Moles, 1958). The results point out that little specialized data helps significantly the simplification.

Références

CARDON R. & GRABAR N. (2019). Parallel sentence retrieval from comparable corpora for biomedical text simplification. In *Proceedings of the International Conference on Recent Advances*

in Natural Language Processing (RANLP 2019), p. 168–177, Varna, Bulgaria : INCOMA Ltd. DOI : 10.26615/978-954-452-056-4_020.

CARDON R. & GRABAR N. (2020). French biomedical text simplification : When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, Barcelona, Spain (online) : Association for Computational Linguistics.

KANDEL L. & MOLES A. (1958). Application de l'indice de flesch à la langue française. *The Journal of Educational Research*, **21**, 283–287.

KLEIN G., KIM Y., DENG Y., SENELLART J. & RUSH A. M. (2017). OpenNMT : Open-source toolkit for neural machine translation. In *Proc. ACL*. DOI : 10.18653/v1/P17-4012.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

XU W., NAPOLES C., PAVLICK E., CHEN Q. & CALLISON-BURCH C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, **4**, 401–415. DOI : 10.1162/tacl_a_00107.

ZHANG X. & LAPATA M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 584–594, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1062.