Simplification automatique des textes scientifiques

Julien Jouan

Université de Bretagne Occidentale, 20 Rue Duquesne, 29200, Brest, France

Abstract

Ce document est un résumé de la présentation que j'effectuerais pour Mots/Machines. Elle concerne la simplification automatique des textes scientifiques grâce à la création d'un corpus sur le thème de l'écologie. Cette simplification automatique est possible grâce au logiciel Jurassic disponible sur le site internet Studio AI21. J'expliquerais les démarches suivies pour réaliser mon étude, et présenterais une partie de mes résultats dans l'optique de recevoir des avis et des suggestions.

Keywords

Simplification, Automatique, Jurassic, Écologie, Biologie, Langues, Linguistique, Corpus, Scientifique, CEUR-WS

1. Introduction

Ma participation à cette journée Mots/Machines est directement liée à mon mémoire. J'ai décidé de consacrer ce mémoire à l'étude de la simplification automatique de textes. Dans ce cas précis, il est question de textes scientifiques liés à l'écologie. Les textes que j'étudie sont tous issus du site internet Papier-Mâché. Ce site répertorie des articles scientifiques vulgarisés provenant tous d'articles et d'études scientifiques rédigés par des scientifiques. J'ai donc créé un corpus commun composé d'environ 35000 mots, corpus qui lui est divisé en deux sous-corpus, à savoir les textes vulgarisés d'un côté et les textes scientifiques de l'autre. La création de ce corpus a plusieurs objectifs :

- aligner manuellement les phrases des deux sous-corpus scientifiques et vulgarisés;
- repérer les différences notables, ainsi que les procédés et styles d'écriture des sous-corpus;

Après avoir réalisé ce travail de repérage autour du corpus et mis en parallèle les phrases, la phase de simplification automatique entre en jeu. Cette simplification automatique est réalisée à partir du site internet Studio AI21. Ce site internet exploite des données à l'aide d'un modèle d'intelligence artificielle nommé Jurassic-1. Ce modèle regorge d'utilisations diverses et variées, et l'une des applications de Jurassic-1 est donc la simplification automatique de textes, ce qui nous intéresse ici particulièrement. Il est possible, à partir de ce logiciel, de générer entre autre des définitions à l'image d'un dictionnaire, ou encore de simplifier des phrases utilisant un jargon d'un domaine précis. Il est important de préciser que le logiciel emmagasine les données : plus les données sont conséquentes et plus le logiciel devient performant. Il est également possible de régler le degré de simplification auquel nous souhaitons que le logiciel travaille. Sur une échelle de 0 à 1, plus la « température » est proche de 1 et plus le logiciel aura tendance à

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

formuler des réponses variées et créatives, mais peut parfois s'éloigner du sujet et d'une phrase correcte scientifiquement. Dans un premier temps, les phrases et définitions de mon sous-corpus d'articles scientifiques sont passées dans ce logiciel afin d'observer les simplifications automatiques. Ensuite, le principe est de rassembler ces phrases et de les comparer avec les phrases vulgarisées rédigées par les rédacteurs de Papier-Mâché. Il y a plusieurs objectifs dans ce travail :

- essayer d'observer les différences entre la simplification manuelle et automatique ;
- repérer des potentielles tendances concernant la simplification automatique ;
- voir si la simplification automatique de textes scientifique (ou même plus globalement) est fiable, en comparant avec d'autres études menées préalablement ;
- comprendre les points forts et les points faibles, problèmes et faiblesses de la traduction automatisée par ordinateur.

2. Bibliographie

- Camille Lemonnier, Naviguer quand on n'a jamais navigué : mieux comprendre le comportement en mer des manchots royaux grâce aux données satellites, 2021, Papier-Mâché, <https://papiermachesciences.org/2021/10/19/naviguer-quand-on-na-jamais-navigue-mieux-comprendre-le-comportement-en-mer-des-manchots-royaux-grace-aux-donnees-satellites/?v=C>.
- Benjamin Dupuis, Les albatros, nos meilleurs alliés contre la pêche illégale, 2021, Papier-Mâché, <https://papiermachesciences.org/2021/08/01/les-albatros-nos-meilleurs-allies-contre-la-peche-illegale/>.
- Moïra Courseaux, L'eau se dégaze ! En quête de la recette de la production de méthane aquatique, 2021, Papier-Mâché, <https://papiermachesciences.org/2021/02/11/leau-se-degaze-en-quete-de-la-recette-de-la-production-de-methane-aquatique/?v=C>.
- Anne-Sophie Masson, Les nématodes : des animaux si petits et si abondants sur terre, 2021, Papier-Mâché, <https://papiermachesciences.org/2020/11/06/les-nematodes-des-animaux-si-petits-et-si-abondants-sur-terre/>.
- F. Orgeret et al., Exploration during early life: distribution, habitat and orientation preferences in juvenile king penguins, 2019, Movement Ecology, <https://movementecologyjournal.biomedcentral.com/articles/10.1186/s40462-019-0175-3>.
- Henri Weimerskirch et al., Ocean sentinel albatrosses locate illegal vessels and provide the first estimate of the extent of nondeclared fishing, 2020, PNAS, <https://www.pnas.org/doi/full/10.1073/pnas.1915499117>.
- Charlotte Grasset et al., The transformation of macrophyte-derived organic matter to methane relates to plant water and nutrient contents, 2019, Limnology and Oceanography, <https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lno.11148>.