Combler les lacunes de Wikipédia : tirer parti de la génération de texte pour améliorer la couverture encyclopédique des groupes sous-représentés

Simon Mille[1], **Massimiliano Pronesti**[1,2], **Craig Thomson**[1], **Michela Lorandi**[1], **Sophie Fitzpatrick**[3], **Rudali Huidrom**[1], **Mohammed Sabry**[1], **Amy O'Riordan**[3], **Anya Belz**[1]
[1]ADAPT, Dublin City University, [2]IBM Research, [3]Wikimedia Community Ireland
simon.mille@adaptcentre.ie

RÉSUMÉ

Wikipédia a des lacunes systématiques dans sa couverture des langues peu dotées ainsi que des groupes sous-représentés (par exemple, les femmes). Cet article présente un nouvel outil pour soutenir les efforts visant à combler ces lacunes en générant automatiquement des débuts d'articles en anglais, français et irlandais, et en facilitant la post-édition et la mise en ligne sur Wikipédia. Un générateur basé sur des règles et un LLM sont utilisés pour générer deux articles alternatifs à partir de graphes de connaissances DBpedia ou Wikidata sélectionnés par l'utilisateur, permettant à l'article généré via LLM, souvent plus fluide mais plus sujet aux erreurs, d'être vérifié en termes de contenu par rapport à l'article généré par des règles, plus fiable, mais moins fluide. Le code de l'outil est disponible sur https://github.com/dcu-nlg/wiki-gen-demo et il est actuellement déployé sur http://ec2-18-224-151-90.us-east-2.compute.amazonaws.com:3000/.

ABSTRACT

Filling Gaps in Wikipedia : Leveraging Data-to-Text Generation to Improve Encyclopedic Coverage of Underrepresented Groups

Wikipedia is known to have systematic gaps in its coverage that correspond to under-resourced languages as well as underrepresented groups. This paper presents a new tool to support efforts to fill in these gaps by automatically generating English, French and Irish draft articles and facilitating post-editing and uploading to Wikipedia. A rule-based generator and an input-constrained LLM are used to generate two alternative articles from DBpedia or Wikidata knowledge graphs selected by the user, enabling the often more fluent, but more error-prone, LLM-generated article to be content-checked against the more reliable, but less fluent, rule-generated article. The code for the tool is available at https://github.com/dcu-nlg/wiki-gen-demo, and it is currently deployed at http://ec2-18-224-151-90.us-east-2.compute.amazonaws.com:3000/.

MOTS-CLÉS : TAL, Génération de texte, Wikipedia, Multilinguisme.

KEYWORDS: NLP, Natural language generation, Wikipedia, Multilinguality.

ARTICLE : **Accepté à** 17th International Natural Language Generation Conference [1].

1. https://aclanthology.org/2024.inlg-demos.6/