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

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

Correspondence: simon.mille@adaptcentre.ie

Abstract

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 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, enabling the often more fluent, but error-prone, LLM-generated article to be content-checked against the more reliable, but less fluent, rule-generated article.

1 Introduction

Knowledge equity is one of two strategic directions in Wikimedia's 2030 Movement Strategy (Wikimedia Movement, 2017). Systematic content gaps identified in the Wikimedia Knowledge Gap Taxonomy[1] relate e.g. to gender, recency, geography and language. For instance, women and non-binary people make up less than 20% of biographies on Wikipedia.[2] Addressing such gaps is important for equity and knowledge completeness, but despite increasing awareness, existing social, political and technical barriers still make it difficult to motivate and/or enable Wikipedia editors to fill them in.

Natural Language Generation (NLG) has held the ambition to help address such gaps for some time (Sauper and Barzilay, 2009; Banerjee and Mitra, 2016; Liu et al., 2018; Fan and Gardent, 2022; Shao et al., 2024), and recent work has shown it to be a valid approach (Kaffee et al., 2022). However, data-to-text NLG has a high knowledge threshold, and the more recent LLM-based NLG systems require substantial cost and energy.[3]

With the tool we report in this paper we aim to address the above issues, targeting underresourced languages and the knowledge threshold in particular. Our Wikipedia Gap Filler tool generates a draft article in Irish or English from an entity name, making it far easier for users to create texts about given entities that can be used as a starting point for a new Wikipedia page.

The contributions of this paper are: (i) the implementation and release of a tool for generating and editing text snippets on a queried entity, with a human-friendly user interface that integrates all the components of the system; (ii) the implementation and release of code for retrieving information about queried entities on DBpedia and Wikidata. The tool and source code can be found on GitHub.[4]

2 Background and Tool Overview

Kaffee et al. (2022) provided first indications that using NLG is a good strategy for filling in Wikipedia knowledge gaps. In particular they concluded that (i) Machine Translation is not adequate to create new Wikipedia pages, due to the cultural differences between the communities that speak different languages (i.e. each community has their specific points of interest; in addition, source text is not always available); (ii) providing Wikipedia editors with even just a starting sentence as in Kaffee et al.'s experiments can have a real impact on page editing; (iii) text is judged more useful than tables with raw data such as article stubs (Kaffee, 2016); and (iv) the main limitation of using NLG for creating text snippets is the lack of factual accuracy of the produced text.

The Wikipedia Gap Filler tool builds upon these ideas and goes beyond, generating the first sentence that can then be expanded upon to form a more extensive draft articles generated from a user-selected

[1] `https://meta.wikimedia.org/wiki/Research:Knowledge_Gaps_Index/Taxonomy`

[2] `https://en.wikipedia.org/wiki/Help:Mapping_content_gaps_on_Wikimedia`

[3] https://www.theguardian.com/commentisfree/article/2024/may/30/ugly-truth-ai-chatgpt-guzzling-resources-environment

[4] `https://github.com/nlgcat/webnlg_demo`

Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations, pages 16–19
September 23–27, 2024. ©2024 Association for Computational Linguistics

set of knowledge triples. Moreover, the tool addresses issues around factual accuracy by generating two texts in parallel, one using a rule-based generator and the other using an input-constrained LLM, enabling the more fluent output from the latter to be content-checked against the more reliable output from the former. Finally, our tool provides users with a selection of entities from known gaps, encouraging them to create pages for these under-represented entities.

Initiatives such as the WikiProject Women in Red (WiR)[5] aim to create Wikipedia content about women's biographies, women's works and women's issues. Lists of women who have no Wikipedia page, i.e. whose name on Wikipedia appears as a red link (hence the name of the initiative), were compiled by the contributors and sorted by category,[6] with 224 categories containing entities for which data is available on Wikidata. That is, for about 400,000 Women in Red, some information is available in the form of triples on (at least) Wikidata, which constitute the input to our Wikipedia Gap Filler tool. We take advantage of this resource to suggest names of WiR to users of our tool.

To use the tool, the user simply enters or selects an entity name and is then offered a list of triples (<entity, property, value>, e.g. <Barack_Obama, birthYear, 1961>) retrieved from DBpedia or Wikidata, from which those to be verbalised in the article can be selected. A wide range of entities can be queried, such as persons, places, buildings, organisations, artistic or intellectual works, fictional characters, vehicles, etc. By using only DBpedia and Wikidata contents as input, the contents of the generated texts are fully traceable, and the length of the text is customisable, by selecting more properties in the input or fewer. The user can request one or more text(s) to be generated, before editing the results in the interface, and from there, create a new Wikipedia page or enrich an existing one.

3 Interface

The Wikipedia Gap Filler app is implemented with a Python Flask back-end and a React JavaScript front-end. The user interface is intended to be very simple to use. First, to input an entity, the user has two alternatives: (i) type in an entity, or (ii) select a suggested entity from the Woman in Red frame. Then, the user can select a grammatical gender (if appropriate), an output language and an NLG system; additionally, the triple source (DBpedia Ontology, Wikipedia Infobox or Wikidata) can be specified. Changing the triple source will return different triples; if a Woman in Red is selected, the default source is Wikidata, for which we know that some triples are available (see Section 2). The user then selects some triples and clicks the Generate button to get the text(s), which can be edited. A button is also available to try and create a new Wikipedia page where the edited text can be pasted.

4 Components

In this section, we briefly describe the main components of our demo: interactive content selection, text generators, and text editing box.

4.1 Interactive content selection

Given an entity, the system first retrieves a list of properties associated with it: (i) the DBpedia/Wikidata SPARQL endpoint URL is defined, (ii) an SPARQLWrapper object is created and the query set, (iii) the query is executed, and (iv) the results are parsed and properties chosen from the DBpedia Ontology,[7] Wikidata,[8] or from a list of properties extracted from Wikipedia Infoboxes but available via DBpedia too,[9] according to the parameters defined in the user query; the Ontology tends to return less triples but of better quality than the Infobox and Wikidata sources. For the moment, the subset of queried properties is limited to about 450, roughly corresponding to the properties in the WebNLG dataset (Gardent et al., 2017; Castro Ferreira et al., 2020), on which the NLG systems we use were developed. A second query is performed on DBpedia to retrieve information about the entities of the triples (e.g. class membership, gender) which is needed by the rule-based system.[10]

4.2 Generators

Once the triples are selected, two systems can be run using the triples as input.

Rule-based pipeline. As a rule-based system, we use the FORGe multilingual generator presented in (Mille et al., 2023), which covers generation

[5] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

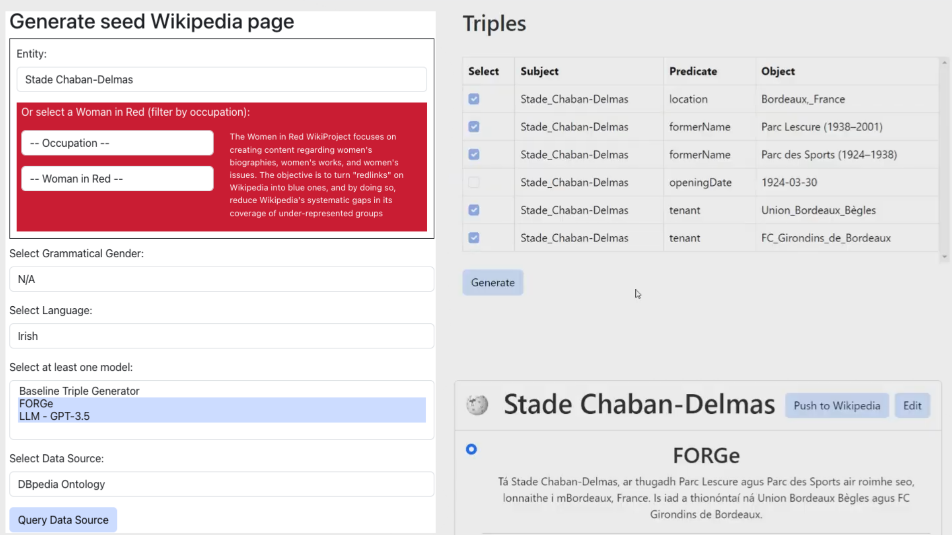
[6] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red/Redlist_index

[7] http://dbpedia.org/ontology/

[8] http://www.wikidata.org/prop/direct/

[9] http://dbpedia.org/property/

[10] The *Select Grammatical Gender* field of our interface overwrites what is found in the query.



Generate seed Wikipedia page

Entity: Stade Chaban-Delmas

Or select a Woman in Red (filter by occupation):

-- Occupation --

-- Woman in Red --

The Women in Red WikiProject focuses on creating content regarding women's biographies, women's works, and women's issues. The objective is to turn "redlinks" on Wikipedia into blue ones, and by doing so, reduce Wikipedia's systematic gaps in its coverage of under-represented groups

Select Grammatical Gender: N/A

Select Language: Irish

Select at least one model:
Baseline Triple Generator
FORGe
LLM - GPT-3.5

Select Data Source: DBpedia Ontology

Triples

Select	Subject	Predicate	Object
☑	Stade_Chaban-Delmas	location	Bordeaux,_France
☑	Stade_Chaban-Delmas	formerName	Parc Lescure (1938–2001)
☑	Stade_Chaban-Delmas	formerName	Parc des Sports (1924–1938)
□	Stade_Chaban-Delmas	openingDate	1924-03-30
☑	Stade_Chaban-Delmas	tenant	Union_Bordeaux_Bègles
☑	Stade_Chaban-Delmas	tenant	FC_Girondins_de_Bordeaux

Stade Chaban-Delmas

○ FORGe

Tá Stade Chaban-Delmas, ar thugadh Parc Lescure agus Parc des Sports air roimhe seo, lonnaithe i mBordeaux, France. Is iad a thionóntaí ná Union Bordeaux Bègles agus FC Girondins de Bordeaux.

Figure 1: Screenshot of the GUI

in several languages including Irish and English. FORGe is implemented as a pipeline of components that perform subtasks such as text planning, lexicalisation, sentence structuring and surface realisation. For Irish morphology, the pipeline uses the finite-state transducers of the Irish NLP tools suite (Dhonnchadha et al., 2003). FORGe is an all-around generator that is not designed to produce specifically Wikipedia-style text, but it can be used without limitations and for free. The output text can be used as part of a seed Wikipedia page or to control the contents delivered by the LLMs.

End-to-end LLM-based generator. The current demo accesses the ChatGPT 3.5 Turbo model[11] via the aiXplain API.[12] We use the Few-Shot In-Context prompt from Lorandi and Belz (2024), which consists of a brief task description followed by two examples showing both input and output and ending with the input. We will add more models such as LLaMa2 70B chat (Touvron et al., 2023), but in general the use of LLMs will be constrained by the availability/cost of the used API.

4.3 Text editor and Wikipedia page creation

The output text is shown in a text box that has an *Edit* mode for the user to modify or combine the texts. Having logged in with their own Wikipedia editor credentials on their browser, users are able to create automatically a new empty Wikipedia page, on which they can paste the selected text.

The present tool is intended to help editors when creating new pages on Wikipedia, but it remains their responsibility to make sure that the uploaded texts respect the detailed Wikipedia edition guidelines (content, style, behaviour, etc.).[13]

5 Limitations

Some of the limitations of our tool are due to the early stage of development that it is in, but others are more general. For the first type, the tool is currently limited to English and Irish texts, and to a subset of about 14% of the DBpedia properties and 1% of the Wikidata properties. This is due to the current coverage of the rule-based generator.

The more general limitations are two-fold: (i) our tool fully depends on the availability of triples for the queried entity, but there can be very little or no information on DBpedia and Wikidata for some under-represented entities; and (ii) despite the help that NLG and LLMs can bring when creating new pages, there still are challenges and potential issues such as the introduction of societal biases and factual errors (Fan and Gardent, 2022), the attribution of the generated contents (Singh et al., 2024), or the actual impact of these technologies on Wikipedia edition (Reeves et al., 2024). Finally, the present tool is not intended to be used in an unsupervised manner, and the potential users are expected to carefully check that the contents they eventually upload to Wikipedia are correct and conform to the Wikipedia guidelines (see Section 4.3).

[11] https://platform.openai.com/docs/models/gpt-3-5-turbo

[12] https://aixplain.com/

[13] https://en.wikipedia.org/wiki/Wikipedia:List_of_guidelines

Acknowledgments

Mille's contribution was funded by the European Union under the Marie Skłodowska-Curie grant agreement No 101062572 (M-FleNS).

References

Siddhartha Banerjee and Prasenjit Mitra. 2016. Wikiwrite: Generating wikipedia articles automatically. In *IJCAI*, pages 2740–2746.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Elaine Uí Dhonnchadha, Caoilfhionn Nic Pháidín, and Josef Van Genabith. 2003. Design, implementation and evaluation of an inflectional morphology finite state transducer for Irish. *Machine Translation*, 18:173–193.

Angela Fan and Claire Gardent. 2022. Generating full length wikipedia biographies: The impact of gender bias on the retrieval-based generation of women biographies. *arXiv preprint arXiv:2204.05879*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Lucie Aimée Kaffee. 2016. *Generating article placeholders from Wikidata for Wikipedia: increasing access to free and open knowledge*. Ph.D. thesis, Hochschule für Technik und Wirtschaft Berlin.

Lucie-Aimée Kaffee, Pavlos Vougiouklis, and Elena Simperl. 2022. Using natural language generation to bootstrap missing wikipedia articles: A human-centric perspective. *Semantic Web*, 13(2):163–194.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Michela Lorandi and Anya Belz. 2024. High-quality data-to-text generation for severely under-resourced languages with out-of-the-box large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1451–1461, St. Julian's, Malta. Association for Computational Linguistics.

Simon Mille, Elaine Uí Dhonnchadha, Lauren Cassidy, Brian Davis, Stamatia Dasiopoulou, and Anya Belz. 2023. Generating Irish text with a flexible plug-and-play architecture. In *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 25–42, Singapore. Association for Computational Linguistics.

Neal Reeves, Wenjie Yin, Elena Simperl, and Miriam Redi. 2024. " the death of wikipedia?"–exploring the impact of chatgpt on wikipedia engagement. *arXiv preprint arXiv:2405.10205*.

Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Aakash Singh, Deepawali Sharma, Abhirup Nandy, and Vivek Kumar Singh. 2024. Towards a large sized curated and annotated corpus for discriminating between human written and ai generated texts: A case study of text sourced from wikipedia and chatgpt. *Natural Language Processing Journal*, 6:100050.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wikimedia Movement. 2017. Wikimedia movement strategy 2017.