Formal systems for people: A safe approach to completion of data and to mitigation of biases against minorities

Marco A. Stranisci* Lorenzo Bafunno* Viviana Bono* Davide Camino*

Rossana Damiano* Emanuele Rovaretto* Andrea Zito*

*University of Turin

Abstract

Our research is aimed at providing a reliable methodology for assessing the presence of social biases and knowledge gaps in the Wikimedia ecosystem. The project combines the power of Language Models with symbolic approaches to extract structured knowledge from Wikipedia biographies, identify biases, and find sources for their mitigation. The prototype of an application based on this pipeline will be released at the end of the project.[1]

Introduction

Despite being one of the most authoritative sources of knowledge, the Wikimedia ecosystem is affected by issues related to discrimination of minorities. Our research project addresses the precise identification, taking into account bias mitigation, based on fine-grained biographical-event detection. Our approach relies on formal systems aimed to represent and manipulate informal knowledge extracted through neural approaches. This way, the outcome is not only a certifiable correct knowledge graph, but also the process to obtain the graph is transparent and modifiable.

The project is organized in two steps:

1. Intrinsic evaluation of biographies. We will gather all English Wikipedia biographies with a corresponding Wikidata page, automatically extract their biographical events and identify forms of underrepresentation and social biases.

2. Extrinsic evaluation. We will create a tool that, given the Wikipedia biography of a person, gathers a number of external documents about the same entity and transforms them into structured knowledge, in order to identify potential mitigation strategies against underrepresentation through extra facts.

Related work

Existing works on bias detection in the Wikimedia ecosystem are affected by two flaws: they rely on coarse-grained NLP analysis (Field *et al.* 2022) and are intrinsic to Wikipedia (Sun *et al.* 2021). Recently we proposed a fine-grained approach to bias detection (Stranisci *et al.* 2023) that can be generalizable also to any document

[1] The repository of the project is available at: https://github.com/DavideCamino/bando_wikimedia

in order to compare Wikipedia with other knowledge bases.

Methods

Actually, our proposal is a pipeline of tools. Given a set of external biographies about a person, we automatically extract triples about them using a classifier. As the second tool in the pipeline, we plan to utilize CDuce (Benzaken *et al.* 2003), which is an XML-oriented functional language with types. CDuce supports the generation and the checking of XML documents, with respect to a given formal specification (i.e., a type). As such, it can support in a certified and traceable manner completions and corrections of the semi-structured information.
The challenge we intend to tackle is twofold:

- to incorporate stochastic information (extracted by the classifier) in the type checking;
- formalize in types otherwise fuzzy concepts such as gender biases. To do so, we plan to involve different communities to collect tags that define certain concepts and include Social Sciences and Psychology experts.

Expected output

The outcome of the pipeline is a knowledge graph-based interface that can be exploited by the editors of Wikipedia to: (i) identify knowledge gaps and social biases in the encyclopedia; (ii) assess new knowledge to be integrated in Wikipedia; (iii) support contributors' discussion on relevant issues related to the representation of minorities. Moreover, the pipeline will support the automatic integration of approved knowledge graphs, relieving the tasks of the moderators' work.

Risks

The first risk of the project is about how contributors will accept to be supported by an AI-enhanced tool for bias discovery. A second risk is the integration of this tool within Wikidata practices. Despite being based on an ontology, Wikidata does not have hard semantic constraints.

Community impact plan

The project is expected to have a crucial impact in the development of a transparent methodology for the evaluation of the quality of knowledge expressed in the Wikimedia ecosystem. Our work will contribute to measuring objective vulnerabilities that are present in Wikipedia biographies. This will ensure: (i) a more informed decision-making process between Wikimedia contributors; and (ii) a continuous support to the identification of content gaps in Wikipedia and Wikidata.

Evaluation

Our team will create a benchmark corpus for an evaluation of performances of each stage of our pipeline: (i) event detection, (ii) generation and checking of structured knowledge. Data and code will be available for experiment replication. Evaluation will be performed by adopting existing metrics widely used in NLP, such as F-1 score. Additionally, a first prototype will be made available for the Wikimedia community to be tested through a UX evaluation module.

Budget

The budget provided for our research is 50,000$ and equally covers personnel costs allocated to the two macro-activities of the project: 22,500 dollars for the implementation of the biographical event extraction pipeline; 22,500 dollars for the development of the validation

tool based on CDuce. The remaining 5,000 dollars will be used to support the dissemination strategy of the project.

Prior contributions

The team has gained relevant experience in the analysis of social biases and underrepresentation in the Wikimedia ecosystem. Recent work has been done to develop a classifier for biographical event detection (Stranisci, Damiano *et al.* 2023) that has been adopted to investigate the most stereotypical correlation between non-Western writers and their biographical events in English Wikipedia pages. Another work (Stranisci, Bernasconi *et al.* 2023) explored strategies to investigate and reduce the underrepresentation of non-Western writers in Wikipedia and Wikidata. Scientific outcomes of these researches will be generalized within the project to people working in different fields.

References

Benzaken, V., Castagna, G., & Frisch, A. (2003). CDuce: an XML-centric general-purpose language. *ACM SIGPLAN Notices*, *38*(9), 51-63.

Field, A., Park, C. Y., Lin, K. Z., & Tsvetkov, Y. (2022, April). Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022* (pp. 2624-2635).

Stranisci, M. A., Bernasconi, E., Patti, V., Ferilli, S., Ceriani, M., & Damiano, R. (2023, October). The World Literature Knowledge Graph. In *International Semantic Web Conference* (pp. 435-452). Cham: Springer Nature Switzerland.

Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele Radicioni, and Tommaso Caselli. 2023. WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto, Canada. Association for Computational Linguistics.

Sun, J., & Peng, N. (2021, August). Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 350-360).