Heterogeneous affiliation metadata enrichment with AffilGood

Nicolau Duran-Silva [1,3], Pablo Accuosto [1], Berta Grimau [1], Nicandro Bovenzi [1], Piotr Przybyła [3,4], Horacio Saggion [3]

[1] *SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain*
[2] *Research Division of SIRIS Academic, Barcelona, Spain*
[3] *LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain*
[4] *Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland*

Purpose

Availability and access to research outcomes is increasingly becoming less of a challenge within the context of open data and open science (Fuster-Martí et al., 2020). This progress is largely driven by the rise and growth of open scholarly knowledge graphs (Manghi et al., 2019; Priem et al., 2022; Kinney et al., 2023), open research information providers (Wilkinson, 2010; Hendricks et al., 2020; Peroni & Shotton, 2020), and the growing trend of governments and public agencies releasing their research and innovation (R&I) policy data (Fuster et al., 2023). Despite these advancements, effectively curating metadata for large open databases remains a significant challenge. In OpenAlex (Priem et al., 2022), for example, there are more than 8 million different raw affiliation strings in publications produced in 2023, with 72% of those not present in publications from the previous five years. To process this amount of texts is not feasible by humans, due to the enormous and growing number of new organizations and signature variants, therefore there is a strong need for building reliable automated methods.

The task of automatically identifying organizations in author-provided affiliation strings and linking them to unique identifiers from global registries, such as the Research Organization Registry (ROR) or Wikidata, is known as institution name disambiguation or affiliation normalization. Linking scientific works to a regularly-updated human-curated registry of organizations is crucial for addressing organization changes over time, including institutional mergers and splits, as well as evolving naming conventions (Purnell, 2022). Accurate normalization of institutions is vital for research evaluation (Huang et al., 2014) and essential for analyzing scientific production trends, particularly within an open science context (L'Hôte & Jeangirard, 2021). Furthermore, research assessment may be affected by wrong attribution of publications to institutions (Donner et al., 2020; Purnell, 2022).

To precisely attribute publications to institutions can be challenging due to the fact that organizations are frequently mentioned in diverse and unstructured manners, employing various patterns, languages, and abbreviations. In addition, automatically extracted affiliation strings often include noise, irrelevant information, or typographical errors. Affiliations can also refer to different institutional levels, such as departments and collaborative institutions, adding complexity to the task.

In the context of open science (Rafols, 2024), ensuring multilingualism and inclusivity is key to equity, as research now extends beyond traditional higher education and research institutions (HERI) to companies, public administrations, and non-profits. Accurately identifying these contributors is essential for a comprehensive global research landscape.

Otherwise, the scientific contributions of non-traditional actors and diverse regions are at risk of being underrepresented, affecting policy decisions and research assessments.

This work furthers the experiments presented in the 2024 SDP Workshop (Duran-Silva et al., 2024). We contribute with new evaluation datasets, new tools and models and more accessible solutions, such as a Python package and a set of open models for extracting information of raw affiliation strings to address a wide range of use cases and needs.

Methods

We propose a set of tools to support a multistep affiliation normalization pipeline, as well as different use cases, composed by the following steps:

- Affiliation-span identification model

- Affiliation entity recognition model

- Geographical metadata enrichment/normalization

- Entity linking module

We have adapted two language models to the unique ""language"" of affiliations, which has a distinct structure and grammar compared to regular natural language: an English one and a multilingual one. These adapted models are used as a base for more robust information extraction models.

We tackle the obstacle posed by the limited availability of annotated data for these tasks - in particular, for complex and/or multilingual cases - by compiling (creating/or curating and refining) new datasets with which to train and/or evaluate our modules individually, as well as the whole pipeline.

Affiliation span identification task is aimed at extracting and cleaning affiliation strings when there is noise and/or when there are multiple affiliation strings in the same signature. We annotate a dataset containing 2,027 raw affiliation strings, and train English and multilingual models. Their performance is reported in Table 2.

Affiliation entity recognition in affiliation strings not only enables more effective linking with external organization registries, but it can also play an essential role in the geolocation of organizations and can also contribute to identifying organizations and their position in an institutional hierarchy - especially for those not listed in external databases. Information automatically extracted by means of a NER model can also facilitate the construction of knowledge graphs, and support the development of manually curated registries. After analyzing hundreds of affiliations from multiple countries and languages, we defined seven entity types: SUB-ORGANIZATION, ORGANIZATION, CITY, COUNTRY, ADDRESS, POSTCODE, REGION. We annotate a dataset containing 5,266 raw affiliation strings, and train English and multilingual models for this task.

The geographical metadata enrichment combines the geographical entities identified by the NER model, to query the geocoding service of OpenStreetMap, to gather additional geographic information as well as a normalized version of the country name. This can increase the accuracy of the affiliation normalization process by providing more precise geographical metadata. By linking identified geographical entities such as cities and countries to the OpenStreetMap geocoding service, the methodology enhances the ability to resolve ambiguities and inconsistencies in affiliation strings.

Seven evaluation datasets were developed and/or curated for linking raw affiliation strings to the ROR identifiers of institutions mentioned in them. The datasets are designed to provide a rich coverage of examples with different levels of difficulties.

Results

We report results obtained evaluating available systems and our proposed modules on the seven datasets, as well as illustrative examples of the use of geographical metadata enrichment from raw affiliation strings.

Modular tool for institution name disambiguation

We perform an evaluation of other available methods, those available for OpenAlex, Semantic Scholar, and OpenAIRE, and Elasticsearch, on the 7 datasets we developed, as Table 7 presents. As well as the incorporation of the NER module in existing methods, and the exploration of LLMs for candidate selection from a subset of candidate organizations. However, the tool provided allows choosing among different models for selecting candidates.

Value

We introduce AffilGood, an open-source solution designed to enhance the accuracy of institution name disambiguation. This framework includes a Python package, which features new evaluation datasets and models, and a collection of open models for extracting information from raw affiliation strings to support a wide variety of use cases, needs and data sources. Key challenges addressed comprise uniquely identifying organizations in multilingual and complex affiliation strings, and dealing with noisy, ambiguous, and incomplete data.

The Research Organization Registry (ROR) is widely used for normalization, but has limitations for several use cases. To address these gaps, AffilGood provides tools to integrate additional external catalogs and registries of institutions, offering broader applicability across use cases. Furthermore, the new framework includes geographical metadata enrichment that supports the localization of organizations at different administrative levels, improving the precision of territorial analysis.

References

Donner, P., Rimmert, C., & van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. Quantitative Science Studies, 1(1), 150-170.

Duran-Silva, N., Accuosto, P., Przybyła, P., & Saggion, H. (2024, August). AffilGood: Building reliable institution name disambiguation tools to improve scientific literature analysis. In Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024) (pp. 135-144).

Fuster Marti, E., Marinelli, E., Plaud, S., Quinquilla, A., & Massucci, F. (2020). Open Data, Open Science & Open Innovation for Smart Specialisation Monitoring (No. JRC119687). Joint Research Centre.

Fuster, E., Fernández, T., Carretero, H., Duran-Silva, N., Guixé, R., Pujol, J., ... & Romagosa, M. (2023). Towards building a monitoring platform for a challenge-oriented smart specialisation with RIS3-MCAT. STI 2023 Conference Proceedings.

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies, 1(1), 414-427.

Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). Institution name disambiguation for research assessment. Scientometrics, 99, 823-838.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., ... & Weld, D. S. (2023). The semantic scholar open data platform. arXiv preprint arXiv:2301.10140.

L'Hôte, A., & Jeangirard, E. (2021). Using Elasticsearch for entity recognition in affiliation disambiguation. arXiv preprint arXiv:2110.01958.

Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., ... & De Bonis, M. (2019). The OpenAIRE research graph data model. Zenodo.

Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies, 1(1), 428-444.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv. arXiv preprint arXiv:2205.01833.

Purnell, P. J. (2022). The prevalence and impact of university affiliation discrepancies between four bibliographic databases—Scopus, Web of Science, Dimensions, and Microsoft Academic. Quantitative Science Studies, 3(1), 99-121.

Rafols, I. (2024). Rethinking Open Science: shifting from access to connections to care for equity and inclusion. OSF.

Wilkinson, M. (2010). Datacite: The International Data Citation Initiative: Datasets Programme.