AffilGood: Building reliable institution name disambiguation tools to improve scientific literature analysis

Nicolau Duran-Silva[1,2], **Pablo Accuosto**[1], **Piotr Przybyła**[2,3], **Horacio Saggion**[2],
[1]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain,
[2]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain,
[3]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract

The accurate attribution of scientific works to research organizations is hindered by the lack of openly available manually annotated data–in particular when multilingual and complex affiliation strings are considered. The AffilGood framework introduced in this paper addresses this gap. We identify three sub-tasks relevant for institution name disambiguation and make available annotated datasets and tools aimed at each of them, including i) a dataset annotated with affiliation spans in noisy automatically-extracted strings; ii) a dataset annotated with named entities for the identification of organizations and their locations; iii) seven datasets annotated with the Research Organization Registry (ROR) identifiers for the evaluation of entity-linking systems. In addition, we describe, evaluate and make available newly developed tools that use these datasets to provide solutions for each of the identified sub-tasks. Our results confirm the value of the developed resources and methods in addressing key challenges in institution name disambiguation.

1 Introduction

The availability and access to research outcomes are gradually becoming less of an issue in an open data and open science context (Fuster et al., 2020). This is due, in part, to the emergence and expansion 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), and the increasing trend of governments and public agencies to release their research and innovation (R&I) policy data (Fuster et al., 2023). However, effective metadata curation of large open databases remains a significant challenge. A persistent issue is how to uniquely identify institutions involved in research from unstructured affiliation strings in scientific publications, where organizations are mentioned in non-standard formats–even when authors are encouraged to use official institution name signatures when publishing (Purnell, 2022).[1]

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)[2] or Wikidata–is known as *institution name disambiguation* or *affiliation normalization*. Linking scientific works to a regularly-updated human-curated registries 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 and Jeangirard, 2021). Furthermore, research assessment may be affected by wrong attribution of publications to institutions (Purnell, 2022; Donner et al., 2020). 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,[3] 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. Illustrative examples of these challenges are presented in Table 1. Relevant examples of ambiguity are presented by Huang et al. (2014).

Despite several tools and methods having been proposed to tackle different subtasks, as described in

[1] 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.

[2] ROR is a community-led registry of open persistent identifiers for research organizations available at `https://ror.org`

[3] From a sample of 50K affiliations from OpenAlex, we estimate that 23% of the strings are not in English.

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 135–144
August 16, 2024 ©2024 Association for Computational Linguistics

Challenge	Example
Non-English input	*Instytut Badań Literackich PAN, Zespół Badań nad Literaturą Zagłady*
Lack of punctuation	*IUM - INSEEC Research Center Strategy & Management Department International University of Monaco Monte-Carlo Monaco*
Overuse of acronyms	*IMSIA, ENSTA Paris, CNRS, CEA, Inst. Polytechnique de Paris, Palaiseau, France*
Diverse locations	*Deanery of Biomedical Sciences, University of Edinburgh, and TW2Informatics Ltd, Gothenburg, 42166, Sweden.*
Multiple organizations	*Virgen del Rocío University Hospital, Network Centre for Biomedical Research in Mental Health (CIBERSAM), Institute of Biomedicine of Seville (IBiS), University of Seville, First-episode Psychosis Research Network of Andalusia (Red PEPSur).*
Noisy content	*#N##TAB##TAB# California Institute of Technology#N##TAB# (S.W., C.H., J.H.T., M.F., M.G.L., D.E.H., T.E., L.K.)*
Irrelevant content	*emeritus professor of child and adolescent psychiatry at the Academic Medical Center in Amsterdam. Since his retirement (2009), he has been practicing the profession from behind the writing table.*
Missing information	*Faculty of Computer Science*

Table 1: Challenges in affiliation normalization.

Section 2, the problem of affiliation normalization is far from being solved, since most existing methods do not address most of the challenges presented in Table 1 and, there is no standard evaluation data available for the task (Donner et al., 2020; L'Hôte and Jeangirard, 2021).

Our contributions include: i) Providing training and evaluation datasets for affiliation span identification and named entity recognition (NER) subtasks, along with seven entity-linking validation datasets with varying difficulty levels (including multiple institutions, diverse formats, and several languages), and the disambiguated institutions with their ROR identifiers; ii) fine-tuning and evaluating new NER models based on RoBERTa and XLM-RoBERTa for identifying entities in affiliation strings (organizations and their locations); iii) using the entities predicted by our NER models as input for a two-step entity linking module that first retrieves candidate ROR identifiers via an information retrieval system, i.e. Elasticsearch, and then re-ranking them using a quantized generative model; iv) using our datasets to compare our modules' and the full pipeline's performance against existing methods; v) making datasets, code, and models available to the research community.[4] To our knowledge, this is the first effort to provide open, manually annotated data for institution name disambiguation in scientific papers and R&I projects.

2 Related work

Various approaches have been proposed to tackle the disambiguation of institution names, including knowledge- and rule-based approaches, name matching and search techniques, supervised learning, and clustering. Rule-based methods (Jonnalagadda and Topham, 2011; Donner et al., 2020) require great effort to develop and maintain rules and often depend on external sources such as thesauri containing organization names and variants. Although attempts have been made to automate the creation of rules (Shao et al., 2020) it is still challenging to extend rule-based methods to a global scale and to adapt them to growing and changing data. Clustering methods have been proposed in combination of rules to normalise and group name variants (Jonnalagadda and Topham, 2011; Cuxac et al., 2012). However, additional manual cleaning steps are often necessary to obtain reliable results.

L'Hôte and Jeangirard (2021) introduce *affiliation-matcher*, a tool aimed at linking affiliations to different registries of organizations, including ROR and Sirene,[5] using the Elasticsearch search engine[6] to match organization names and locations in affiliations and the destination registries. This approach, however, presents some limitations when information available in affiliation strings is noisy or with significant variations with respect to the data indexed in the registries. The *OpenAlex Institution Parsing*[7] tool also links affiliations strings to ROR institutions, and is the affiliation disambiguation system used in the OpenAlex database (Priem et al., 2022). The authors of this system model the normalization task as an extreme multi-label text classification problem: they propose to use two DistilBERT sequence classification models to predict ROR identifiers trained on synthetic affiliation strings obtained from the OpenAlex and based on different affiliation templates (including, for instance, organization names, aliases, acronyms, cities, regions, and countries). *AffRo*,[8] the affiliation matching algorithm used in OpenAIRE (Manghi et al., 2019), pre-processes affiliation strings in order to identify relevant segments in them and then employs cosine similarity for matching organization names with ROR entries, considering different similarity thresholds for universities and non-university organizations.

[4] Our code, models, and datasets are available at https://github.com/sirisacademic/affilgood

[5] Sirene is the "Système National d'identification et Du Répertoire Des Entreprises et de Leurs établissements" available at https://www.sirene.fr/

[6] https://www.elastic.co

[7] https://github.com/ourresearch/openalex-institution-parsing

[8] https://github.com/openaire/affro

S2AFF,[9] the affiliations linker module used in Semantic Scholar (Kinney et al., 2023) implements a three-steps method: a NER model first parses raw affiliation strings to extract main and child organizations as well as address components. The identified entities are used to retrieve a set of candidates from a ROR dump, which are then re-ranked by means of a parwise feature-based model, as described in 4.3.1. The main limitation of the S2AFF linker is that it has not been trained to deal with non-English affiliations. Chen et al. (2023) explore the use of a deep learning model to normalize organization names according to closed scholarly knowledge graphs: they classify parts of the affiliation strings as *institutional data* (first level and second level institutions) and *non-institutional data* (for address components) and then apply an institution matching and merging model that uses word embeddings and a set of manually formulated rules for data transformation and processing. The system also includes a relation extraction model to identify hierarchical institutional relationships.

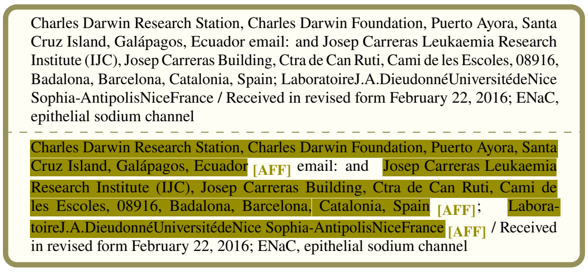
3 Task definition and datasets

In line with previous work in this area (Kinney et al., 2023; Chen et al., 2023), we propose to approach affiliation normalization as a three-step pipeline that includes the following sub-tasks: (1) raw affiliation span identification (3.1), (2) named entity recognition (3.2), and (3) entity linking (3.3). 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.

Addressing the unique identification of organizations in affiliation strings as three related subtasks has the additional advantage of allowing more flexibility, as each of the modules can be used independently in different downstream applications.[10]

3.1 Affiliation span identification

Raw 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. Typically, multiple institutions have been considered to be separated by semicolons. However, other punctuation marks, spaces or *and* connectors are frequently used to separate affiliations (see Fig 1).



Charles Darwin Research Station, Charles Darwin Foundation, Puerto Ayora, Santa Cruz Island, Galápagos, Ecuador email: and Josep Carreras Leukaemia Research Institute (IJC), Josep Carreras Building, Ctra de Can Ruti, Cami de les Escoles, 08916, Badalona, Barcelona, Catalonia, Spain; LaboratoireJ.A.DieudonnéUniversitédeNice Sophia-AntipolisNiceFrance / Received in revised form February 22, 2016; ENaC, epithelial sodium channel

Charles Darwin Research Station, Charles Darwin Foundation, Puerto Ayora, Santa Cruz Island, Galápagos, Ecuador [AFF] email: and Josep Carreras Leukaemia Research Institute (IJC), Josep Carreras Building, Ctra de Can Ruti, Cami de les Escoles, 08916, Badalona, Barcelona, Catalonia, Spain [AFF]; LaboratoireJ.A.DieudonnéUniversitédeNice Sophia-AntipolisNiceFrance [AFF] / Received in revised form February 22, 2016; ENaC, epithelial sodium channel

Figure 1: Example of affiliation span identification task from raw strings.

Dataset creation. We have annotated[11] a dataset containing 2,072 raw affiliation strings obtained from OpenAlex to identify spans containing relevant affiliation data within them. The annotated instances were selected by a stratified random sampling by country, focusing on ensuring diversity in affiliation languages and origins. Additional manually-chosen instances with noisy sequences were included in the annotated data so we could train our model to filter out non-affiliation strings. As shown in Fig. 1, it is frequent that affiliation data automatically extracted from PDF files contain texts that should have been discarded (e.g. email, acknowledgements or part of the contents of the publication). These data can introduce errors in the subsequent steps of the pipeline.[12]

3.2 Named entity recognition

Identifying named entities (organization names, cities, countries) 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 identify 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, as suggested by Chen et al. (2023), and support the development of manually curated registries.

After analyzing hundreds of affiliations from multiple countries and languages, we defined seven en-

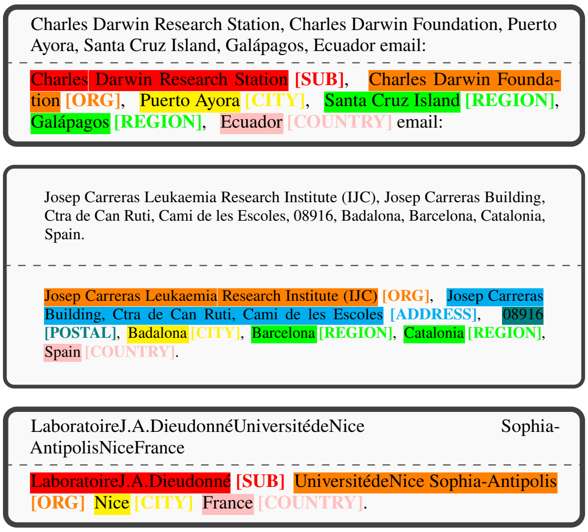
[9] https://github.com/allenai/S2AFF

[10] For instance, the entity linking module can be applied directly to the data sources from where the CORDIS and ETER datasets described in 3.3 are obtained, as they do not require the previous identification of named entities.

[11] For this task a single annotator was in charge of identifying the significant spans in raw strings.

[12] In the example, the string *ENaC* in Figure 1, for instance, could be wrongly assumed to correspond to the acronym of *École Nationale de l'Aviation Civile* in Toulouse, with ROR id https://ror.org/022zdgq74.

tity types: SUB-ORGANISATION, ORGANISATION, CITY, COUNTRY, ADDRESS, POSTAL-CODE, and REGION, detailed in Appendix A.



Charles Darwin Research Station, Charles Darwin Foundation, Puerto Ayora, Santa Cruz Island, Galápagos, Ecuador email:

Charles Darwin Research Station [SUB], Charles Darwin Foundation [ORG], Puerto Ayora [CITY], Santa Cruz Island [REGION], Galápagos [REGION], Ecuador [COUNTRY] email:

Josep Carreras Leukaemia Research Institute (IJC), Josep Carreras Building, Ctra de Can Ruti, Cami de les Escoles, 08916, Badalona, Barcelona, Catalonia, Spain.

Josep Carreras Leukaemia Research Institute (IJC) [ORG], Josep Carreras Building, Ctra de Can Ruti, Cami de les Escoles [ADDRESS], 08916 [POSTAL], Badalona [CITY], Barcelona [REGION], Catalonia [REGION], Spain [COUNTRY].

LaboratoireJ.A.DieudonnéUniversitédeNice Sophia-AntipolisNiceFrance

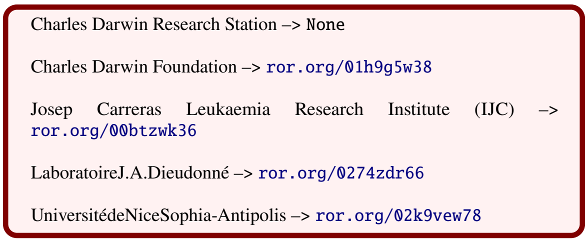
LaboratoireJ.A.Dieudonné [SUB] UniversitédeNice Sophia-Antipolis [ORG] Nice [CITY] France [COUNTRY].

Figure 2: Example of entity recognition task.

Dataset creation. The NER dataset contains 5,266 raw affiliation strings obtained from OpenAlex (a superset of those used in 3.1). It includes multilingual samples from all available countries and geographies to ensure comprehensive coverage and diversity. To enable our model to recognize various affiliation string formats, the dataset includes a wide range of structures, different ways of grouping main and subsidiary institutions and various methods of separating organization names. We also included ill-formed affiliations and those containing errors resulting from automatic extraction from PDF files. Two annotators independently annotated different subsets of samples, as well as an overlapping sample of 100 strings, which we used to compute the inter-annotator agreement at token level, obtaining a macro-averaged $F_1 = 0.962$. Fig. 2 shows two examples of the output of the NER model, and Appendix B shows the frequency of each entity in the dataset.

3.3 Entity linking

As mentioned in 1, unambiguously linking organizations mentioned in affiliation strings and research projects is critical to enable accurate analysis of the scientific production of institutions and the collaboration between them. Fig. 3 shows four different examples of what we would expect the output of a system that links organization names to their ROR identifiers would be.



Charles Darwin Research Station –> None

Charles Darwin Foundation –> ror.org/01h9g5w38

Josep Carreras Leukaemia Research Institute (IJC) –> ror.org/00btzwk36

LaboratoireJ.A.Dieudonné –> ror.org/0274zdr66

UniversitédeNiceSophia-Antipolis –> ror.org/02k9vew78

Figure 3: Example of entity linking task.

Dataset creation. In the context of this work, seven evaluation datasets were developed and/or curated for linking raw affiliation strings to the ROR identifiers of institutions mentioned in them.[13] The datasets are designed to provide a rich coverage of examples with different levels of difficulties, including instances that present several of the challenges described in Table 1. Dataset sizes and sources are presented in Table 2.

Dataset	Source Type	Size	#Orgs/seq.
raw affiliation strings			
Multilingual Affiliations (MA)	Publications	322	0.93
French Affiliations (FA)	Publications	614	3.15
Non Related Multi-orgs (NRMO)	Publications	168	2.82
Mixed Affiliations (S2AFF*)	Publications	635	1.00
pre-segmented entities			
CORDIS	R&I project	3,329	0.83
ETER English (ETERe)	Statistical database	345	0.95
ETER Multilingual (ETERm)	Statistical database	417	0.91

Table 2: Statistics of each entity linking dataset from the raw affiliation strings (top) and pre-segmented entities (bottom). #Orgs/seq.: Avg. number of ROR ids per sequence.

We consider two groups of entity-linking datasets, described below: one group contains raw affiliation strings from scientific publications (MA, FA, NRMO, and S2AFF*), and the other one contains pre-segmented entities from research projects (CORDIS, ETERe, ETERm). The instances in the second group include separate columns for organization names, cities and countries (either country codes or names). The data for the pre-segmented datasets is obtained from UNICS (Gimenez et al., 2018). An example of each dataset is available in Appenix C. Having pre-identified instances allows us to independently evaluate the entity linking component isolating its performance from the errors that can be introduced in the NER module.

For the final pipeline integrating the NER and entity linking modules, we consider the seven datasets. For pre-segmented entities, we join them into a single string with each entity type separated by a colon, the most common format in affiliation strings. Both for the raw affiliation and the pre-segmented datasets, the instances include the manually assigned ROR identifiers with the format:

[13] We used, in all cases, the ROR version 41.

$org\_name_1\{ror\_id_1\}|...|org\_name_n\{ror\_id_n\}$

A strict criterion was applied when assigning gold-standard ROR identifiers: all organizations explicitly mentioned in the affiliation string were included in the list of gold labels, and only those. This means, for example, that assigning an institution the identifier of its parent organization is considered an error, as is an error assigning an institution the identifier of a child organization. Datasets are annotated according to ROR v.41.

Raw affiliation string datasets

- **Multilingual affiliations(MA)**: Manually curated version of a selection of raw affiliations obtained from OpenAlex with `langdetect`[14] for 20 languages[15].
- **French Affiliations (FA)**: Collection of complex French affiliations[16] selected from OpenAlex. Each affiliation was annotated by one annotator and checked/validated by another person.
- **Non-Related Multi-Organizations (NRMO)**: Manually curated version of a selection of raw affiliations obtained from OpenAlex, with non-related organizations in the same string.
- **Mixed Affiliations (S2AFF*)**: Our improved version of the dataset provided by S2AFF,[17] which contains a mix of affiliations from different countries. It was re-annotated according to ROR v.41 by two experts, fixing errors and allowing linking to multiple ROR entries, while the original does not.

Pre-segmented datasets

- **CORDIS**: Organizations involved in EU-funded R&I projects. ROR identifiers were assigned by two annotators. It includes several organizations that are not in ROR. For some organizations, it includes multiple-language versions of their names.
- **ETER English (ETERen)**: English version of organization names in the ETER education statistical registry.[18] ROR identifiers were assigned by two annotators.
- **ETER Multilingual (ETERm)**: Original version of organization names in ETER. ROR identifiers were assigned by two annotators.

4 Methods

As mentioned in 3 we divided the task of linking affiliation strings to ROR identifiers in three subtasks: *affiliation span identification*, *named-entity recognition*, and *entity linking*. In this section we describe the implementation of tools aimed at each of these tasks.

4.1 Adapting language models to raw affiliation strings

For the first two tasks, we fine-tuned two RoBERTa (Liu et al., 2019) and XLM-RoBERTa (Conneau et al., 2019) models for (predominantly) English and multilingual datasets, respectively. Gururangan et al. (2020) show that continuing pre-training language models on task-relevant unlabeled data might contribute to improve the performance of final fine-tuned task-specific models–in particular, in low-resource situations. Considering the fact that the affiliation strings' *grammar* has its own structure, which is different from the one that would be expected to be found in free natural language, we explore whether our affiliation span identification and NER models would benefit from being fine-tuned from models that have been *further pre-trained* on raw affiliation strings for the masked token prediction task.[19] Table 3 reports perplexity on 50k randomly held-out affiliation strings. In what follows, we refer to our adapted models as *AffilRoBERTa*[20] and *AffilXLM*.[21] Hyperparameters used for training are described in Appendix D.

Model	PPL_{base}	PPL_{adapt}
RoBERTa	1.972	1.106
XLM-RoBERTa	1.997	1.101

Table 3: We report masked language modeling loss as perplexity measure (PPL) on 50k randomly sampled held-out raw affiliation strings.

[14] `https://github.com/Mimino666/langdetect`

[15] Languages covered are: fr, de, es, pt, it, id, uk, ca, fa, ko, nl, pl, ro, sw, ru, zh-cn, ar, tr, sv, and ja.

[16] French affiliation strings in scientific papers often feature complex institutional names, including multiple department names, mixed research units, and associated universities, as well as specific terms and abbreviations (CNRS, INSERM, UMR, CHR), special characters, variable formatting, and language mixing.

[17] `https://github.com/allenai/S2AFF`

[18] `https://eter-project.com/`

[19] Further pre-trained on 10 million random raw affiliation strings from OpenAlex.

[20] Available at `https://huggingface.co/SIRIS-Lab/affilgood-AffilRoBERTa`

[21] Available at `https://huggingface.co/SIRIS-Lab/affilgood-AffilXLM`

4.2 Affiliation span and NER models

Both for span identification and entity recognition tasks, we fine-tuned the adapted and base models for token classification with the IOB annotation schema. We trained the models for 25 epochs, using 80% of the dataset for training, 10% for validation and 10% for testing. Hyperparameters used for training are described in Appendix D. The best performing epoch (considering macro-averaged F1 with *strict* matching criteria) was used to select the model.

4.3 Entity linking module

The entity linking module is aimed for associating identified organizations in affiliation strings (pre-identified or obtained from the NER step) with their corresponding ROR identifiers. In developing this module we consider the possibility of affiliation strings containing references to multiple ROR organizations as well as containing none (either because there are no organizations identified in the affiliation string or because the organizations are not included in ROR). Considering the list of entities previously obtained by means of the NER, in this module we first apply a set of heuristics to group named entities into single-organization affiliations.[22] Second, we determine the corresponding ROR identifier for each grouping (or *none* if no reliable match is found). For example, from the raw affiliation string: *Telethon Kids Institute, School of Physiotherapy and Exercise Science, Curtin University, Centre for Child Health, University of Western Australia, Perth, Australia*, we consider the following list of five potential affiliations:

- *Telethon Kids Institute, Perth, Australia*
- *School of Physiotherapy and Exercise Science, Curtin University, Perth, Australia*
- *Curtin University, Perth, Australia*
- *Centre for Child Health, University of Western Australia, Perth, Australia*
- *University of Western Australia, Perth, Australia*

We implement and evaluate three linking methods for each of the groupings obtained as described above: (i) S2AFF linker, (ii) Elasticsearch with no re-ranking, and (iii) Elasticsearch followed by a re-ranking step performed by means of a quantized generative model.

4.3.1 S2AFF linker

The entity linking component of the S2AFF package[23] involves two steps. First, a candidate-retrieval module extracts potential ROR identifiers from a ROR dump[24] by calculating n-gram and token Jaccard similarities between the affiliation string and ROR entries. This list of candidate ROR identifiers and their scores is then passed to a second module, which re-ranks the top N candidates using a trained LightGBM model. The top-ranked candidate from this step is considered the corresponding ROR id. A threshold of 0.20 is set as the minimum score for a reliable match.

4.3.2 Elasticsearch with no re-ranking

Entity linking is often modeled as an information retrieval task using full-text search engines like Apache Solr[25] and Elasticsearch,[26] leveraging their text matching and relevance ranking capabilities (L'Hôte and Jeangirard, 2021). We implemented an Elasticsearch-based entity linker by first indexing the ROR dump in Elasticsearch. During indexing, we enriched organization names and aliases from ROR with Wikidata[27] labels in English and local languages.[28]

The linking component uses queries that consider the organization's name (and its parent if it's a sub-organization) and location (city, region, and/or country). Multiple matching strategies are combined for organization names, including exact, fuzzy, and shingle-based matches. To discard unlikely matches, we set thresholds of 70 for Elasticsearch scores when linking a main organization and 200 for a sub-organization. The thresholds were determined by analyzing the distribution of Elasticsearch scores with a subset of 50 random instances.

4.3.3 Elasticsearch and generative model

As a third alternative for the entity linking component, we included a re-ranking step when the Elasticsearch scores of the top candidate and other retrieved ROR entries were similar. We assumed the top-ranked candidate to be correct if its score was at least three times that of the second candidate. Otherwise, we considered up to five candidates

[22] For instance, when a sub-organization is found, we group it with the first organization to its right, which is considered as the parent.

[23] See https://github.com/allenai/S2AFF.

[24] For our experiments we used the same ROR dump used in the other methods for a reliable comparison of the results.

[25] https://solr.apache.org/

[26] https://www.elastic.co/

[27] https://www.wikidata.org

[28] *E.g.*, for organizations in Spain, we indexed labels in Catalan, English, Euskera, Galego, and Spanish.

and used a generative model to determine the best match for the affiliation string.[29] For this, we used a quantized (4.16 GB) version[30] of Intel's Neural Chat model,[31] a fine-tuned version of the Mistral 7B model (Jiang et al., 2023).[32]

5 Results and discussion

In this section we report the results obtained by each of the three modules (span identification, NER, and entity linking) as well as the integration of the NER and entity linking components, and we compare them to baselines and other existing systems. We have not included the span identification task as part of the pipeline, because as our evaluation datasets contain few examples in which affiliation spans do not cover the whole raw affiliation string, we would obtain limited information to properly conduct an extrinsic evaluation of this modules' performance.

5.1 Raw affiliation span identification

Table 4 shows the results obtained for the affiliation span identification task with the base RoBERTa and XLM-RoBERTa models as well as with the adapted models AffilRoBERTa[33] and AffilXLM[34], obtained with the *further pre-training* strategy described in 4. We report the results obtained when considering both *exact* and *partial* matching criteria for the affiliation spans.[35]

Model	Exact F1	Partial F1
Semicolon split (baseline)	.793	.907
RoBERTa	.929	.981
XLM	.931	.978
AffilRoBERTa	**.938**	**.981**
AffilXLM	.927	.979

Table 4: *Exact* and *partial* F1 scores for raw affiliation span identification.

We observe that there is a gain–of .145 F1-points for the best-performing model with *exact* match–when predicting the affiliation spans by means of the fine-tuned models over the *naive* strategy used as baseline, while the adapted models obtained with the *further pre-training* strategy present a marginal gain with respect of the base models in the best validation epoch.

5.2 NER

Table 5 shows the performance of the NER model considering *strict* matching criteria.[36] We observe that there is a gain F1-points for the best-performing with adapted models. Adapted XLM-RoBERTa[37] achieves the best *strict* F1, and obtains the best result in four of the seven categories, however, adapted RoBERTa[38] also perform competitive results, specially in three of the categories. When we consider the evolution of the validation loss we observe that there is a considerable advantage of the adapted models in the initial fine-tuning epochs, but this distance is reduced in the final epochs. This indicates that the base models also learn to correctly identify the particular structure of the *affiliations' language* over time.

Category	RoBERTa	XLM	AffilRoBERTa	AffilXLM
ALL	.910	.915	.920	**.925**
ORG	.869	.886	.879	**.906**
SUB	.898	.890	**.911**	.892
CITY	.936	.941	.950	**.958**
COUNTRY	.971	.973	**.980**	.970
REGION	.870	.876	.874	**.882**
POSTAL	.975	.975	**.981**	.966
ADDRESS	.804	.811	.794	**.869**

Table 5: NER evaluation (*strict*). F1-score.

5.3 Entity Linking on pre-segmented datasets

Table 6 shows macro-averaged F1 scores obtained for the entity linking task evaluated on the pre-segmented datasets.

Method	CORDIS	ETERe	ETERm
S2AFF$_{Linker}$[39]	.837	.858	.834
Elasticsearch (top-ranked)	.825	.862	.861
Elasticsearch + qLLM	**.884**	**.914**	**.922**

Table 6: Entity linking results. F1 score.

Note that these metrics are not directly comparable to the ones reported for the full pipeline in Table 7 as the latter refers to example-based metrics in a multi-label context while in the results in Table 6 consider a single predicted identifier (or *none*).

[29] The prompt is included as supplementary material.

[30] neural-chat-7B-v3-2-GPTQ

[31] neural-chat-7b-v3-2

[32] Mistral-7B-v0.1

[33] Available at https://huggingface.co/SIRIS-Lab/affilgood-span

[34] Available at https://huggingface.co/SIRIS-Lab/affilgood-span-multilingual

[35] An *exact* match is considered when the boundaries match exactly, while *partial* matches require partial boundary match over the surface strings in the predicted and gold entities (Batista and Upson, 2020).

[36] A *strict* match is considered when both the boundaries and types of the entities match (Segura-Bedmar et al., 2013).

[37] Available at https://huggingface.co/SIRIS-Lab/affilgood-NER-multilingual

[38] Available at https://huggingface.co/SIRIS-Lab/affilgood-NER

[39] https://github.com/allenai/S2AFF

[40] https://github.com/ourresearch/openalex-institution-parsing

[41] https://github.com/allenai/S2AFF

[42] https://github.com/openaire/affro

	Raw affiliation strings				Pre-segmented		
Model	**MA**	**FA**	**NRMO**	**S2AFF***	**CORDIS**	**ETERe**	**ETERm**
ElasticSearch	.545	.407	.470	.515	.751	.855	.847
OpenAlex[40]	.394	.118	.769	**.871**	.648	.859	.852
S2AFF[41]	.546	.367	.617	.785	.649	.668	.720
AffRo[42]	.452	.408	.558	.726	.641	.709	.617
AffilGoodNERm + S2AFF$_{Linker}$	.596	.685	.762	.841	.827	.887	.863
AffilGoodNER + S2AFF$_{Linker}$	.579	.685	.758	.850	.839	.895	.855
AffilGoodNERm + Elastic	.690	.587	.747	.640	.849	.887	.894
AffilGoodNER + Elastic	.649	.610	.755	.648	.855	.893	.881
AffilGoodNERm + Elastic+qLLM	**.710**	.721	**.774**	.790	.881	**.936**	**.916**
AffilGoodNER + Elastic+qLLM	.653	**.747**	.767	.799	**.891**	**.936**	.909

Table 7: Pipeline (NER+EL) results, evaluated by example-based F1-score. AffilGoodNERm correspond to the best-performing fine-tuned NER model with adapted XLM-RoBERTa, and AffilGoodNER, to the best with adapted English RoBERTa. Entities in pre-segmented datasets have concatenated with coma-separator. Disclaimer: we cannot guarantee that any of the baseline systems, such as OpenAlex or S2AFF, used samples from the original version of the S2AFF dataset for training, since it is open.

5.4 NER + Entity Linking pipelines

Table 7 shows example-based F1 scores for four openly available systems and our proposed NER + entity linking pipelines. This includes our two NER models for identifying organization names and locations, combined with the S2AFF entity linking module. Additionally, it includes our proposed system with an Elasticsearch candidate retrieval stage followed by a re-ranking step using a zero-shot quantized generative model, as described in 4.3.3. Appendix 5.3 shows macro-averaged F1 scores obtained for only the entity linking task evaluated on the pre-segmented datasets.

The obtained results show that the multilingual NER model AffilXLM-NER performs better on datasets containing affiliations in multiple languages (MA and ETERm). It can also be observed that the proposed two-step method based on Elasticsearch text matching coupled with a generative re-ranking stage consistently perform well across datasets, indicating their effectiveness in handling diverse affiliation strings and improving linking accuracy. This is particularly clear in the case of complex French affiliations (FA). It can also be observed that both the RoBERTa and XLM-based NER models fine-tuned with our NER dataset contribute to the performance of the whole pipeline. This is made evident when comparing the performance of the full S2AFF pipeline with the performance of the systems obtained when replacing the S2AFF NER with our models and keeping the S2AFF entity linker. The gain is more evident in the case of the difficult French examples, in which our NER models combined with the S2AFF linker obtain a percentage gain of 87% in terms of macro-averaged F1-score over the S2AFF full pipeline.

6 Conclusions

In this work, we have introduced the AffilGood institution name disambiguation framework, including two different datasets for information extraction from raw affiliation strings, and a collection of seven entity linking datasets connecting organizations mentioned in affiliation strings and/or research projects to ROR identifiers. We benchmark our entity linking datasets with openly available institution name disambiguation systems. Finally, we propose a flexible and multilingual multistep pipeline based on a named-entity recognition model and an entity linking module. The obtained results confirm the quality of the contributed datasets and the validity of the proposed systems to address some of the most difficult challenges in the institution name disambiguation task, including noisy and incomplete input data, affiliations in languages other than English and/or with mixed languages, and complex affiliations including diverse types of institutions – companies, universities, hospital, research centers – in different hierarchical levels. To facilitate reproducibility and promote future research in this area we make available all the data and systems developed in the context of this work.

Acknowledgements

Supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya, by Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021), and by Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MCIN/AEI /10.13039/501100011033. This work was co-funded by the EU HORIZON SciLake (Grant Agreement 101058573). The work of P. Przybyła is part of the ERINIA project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders. Neither the European Union nor the granting authority can be held responsible for them.

We thank Nees Jan van Eck and Martijn Visser from the Centre for Science and Technology Studies (CWTS) at Leiden University for providing relevant raw affiliation strings from OpenAlex for compiling the FA and NRMO entity linking evaluation datasets. We thank Justin Barrett from OurResarch for the help in executing the predictions of OpenAlex Institution Parser model, reported in Table 6.

References

David Batista and Matthew Antony Upson. 2020. nervaluate.

Yifei Chen, Xiaoying Li, Aihua Li, Y. Li, Xuemei Yang, Ziluo Lin, Shirui Yu, and Xiaoli Tang. 2023. A deep learning model for the normalization of institution names by multisource literature feature fusion: Algorithm development study. *JMIR Formative Research*, 7.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Pascal Cuxac, Jean-Charles Lamirel, and Valerie Bonvallot. 2012. Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, 97:47 – 58.

Paul Donner, Christine Rimmert, and Nees Jan van Eck. 2020. Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies*, 1(1):150–170.

Enric Fuster, Tatiana Fernández, Hermes Carretero, Nicolau Duran-Silva, Roger Guixé, Josep Pujol-Llatse, Bernardo Rondelli, Guillem Rull, Marta Cortijo, and Montserrat Romagosa. 2023. Towards building a monitoring platform for a challenge-oriented smart specialisation with ris3-mcat. *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*, abs/2401.10900.

Enric Fuster, Elisabetta Marinelli, Sabine Plaud, Arnau Quinquilla, and Francesco Massucci. 2020. Open data, open science & open innovation for smart specialisation monitoring. Technical report, Joint Research Centre (Seville site).

Xavi Gimenez, Alessandro Mosca, Fernando Roda, Bernardo Rondelli, and Guillem Rull. 2018. Unics: The open data platform for research and innovation? In *Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systemsco-located with the 14th International Conference on Semantic Systems (SEMANTiCS 2018)*, volume 2198. CEUR-WS.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. 2020. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427.

Shuiqing Huang, Bo Yang, Sulan Yan, and Ronald Rousseau. 2014. Institution name disambiguation for research assessment. *Scientometrics*, 99:823–838.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Siddhartha R. Jonnalagadda and Philip Topham. 2011. Nemo: Extraction and normalization of organization names from pubmed affiliation strings. *ArXiv*, abs/1107.5743.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Anne L'Hôte and Eric Jeangirard. 2021. Using elasticsearch for entity recognition in affiliation disambiguation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Paolo Manghi, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, and

Pedro Príncipe. 2019. The openaire research graph data model.

Jason Priem, Heather A. Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv*, abs/2205.01833.

Philip J. Purnell. 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.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Zhou Shao, Xiangying Cao, Sha Yuan, and Yongli Wang. 2020. Elad: An entity linking based affiliation disambiguation framework. *IEEE Access*, 8:70519–70526.

Max Wilkinson. 2010. Datacite: The international data citation initiative: Datasets programme.

A Annotation Guidelines

Annotation guidelines are available at `https://github.com/sirisacademic/affilgood`.

B NER dataset description

Table 8 shows number occurrences of each category in the whole dataset, included in the 5266 sequences.

Category	#Occurences
SUB	4,708
ORG	6,200
CITY	4,023
COUNTRY	4,157
REGION	1,024
POSTALCODE	1,424
ADDRESS	788
Total	2,2324

Table 8: Number of occurrences of named entities in our NER dataset.

C Examples of datasets

Table 9 shows a sequence of each of the seven entity linking datasets.

Dataset	Example
MA	*Fakultas Bisnis Universitas Kristen Duta Wacana#N#Jl. Dr. Wihidin Sudiro Husodo 5 - 25, Yogyakarta, 55224*
FA	*Inserm UMR 1011, Department of Cardiovascular Radiology, EGID (European Genomic Institute for Diabetes), université de Lille, Institut Cœur-Poumon, Institut Pasteur de Lille, CHU de Lille, FR3508, 59000 Lille, France*
NRMO	*EMBL Australia Node in Single Molecule Science, School of Medical Sciences, University of New South Wales, Sydney, Australia.*
S2AFF*	*Andrews Univ, Berrien Springs, MI;*
CORDIS	*ZENTRAL-UND LANDESBIBLIOTHEK BERLIN, BERLIN, DE*
ETERe	*University of Applied Sciences Schmalkalden*
ETERm	*Hochschule Schmmalkalden*

Table 9: Examples of sequences in each entity linking datasets.

D Experimental setup

We provide experiental detais of our baseline fine-tuning approaches. For all *futher-pretraining* and *fine-tuning*, we make use of the `huggingface` library. Training was run (using 1x NVIDIA A100 GPU) for all models with hyperparameter defined in Table 10 and Table 11.

Hyperparameter	Value
Learning Rate	2e-5
Learning Rate Decay	Linear
Weight Decay	0.01
Warmup Portion	0.06
Batch Size	128
Num. of steps	25k steps
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.999

Table 10: Hyperparameters for adaptive pre-training to raw affiliation strings.

Hyperparameter	Value
Learning Rate	2e-5
Learning Rate Decay	Linear
Weight Decay	0.01
Batch Size	16
Max. Num. of Epochs	25

Table 11: Hyperparameters for fine-tuning NER.