Use of LLMs to Improve Affiliation Disambiguation in Alexandria3k

Dibyendu Gupta

Supervisor(s): Diomidis Spinellis, Georgios Gousios

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 28, 2024

Name of the student: Dibyendu Gupta
Final project course: CSE3000 Research Project
Thesis committee: Diomidis Spinellis, Georgios Gousios, Koen Langendoen

Abstract

The growth of academic publications, heterogeneity of datasets and the absence of a globally accepted organization identifier introduce the challenge of affiliation disambiguation in bibliographic databases. In this paper, we create a baseline using the currently implemented algorithm for author affiliation linkage in Alexandria3k by comparing it to the ground truth. We aim to explore the usage of LLMs (GPT-4) in the Alexandria3k environment to disambiguate author affiliations. The proposed approach extracts the research organization from textual affiliations provided by researchers through their published works and cross-references the organization across the Research Organization Registry. Our process shows promising results and a significant improvement on the existing algorithm in terms of matching rate and identification of multiple affiliations. We discuss the margin of error in LLM results, limitations of the ground truth, and suggest future research directions.

1 Introduction

The growth of published literature creates a ginormous volume of meta-data that can be used for performing secondary, tertiary and meta-analysis studies [25]. The increasing volume and complexity of academic publications led to the challenge of accurate identification and linking of affiliations in bibliographic databases. The heterogeneity of datasets mentioning affiliations, outdated affiliation storage methods, continual emergence of new research institutions and the lack of globally accepted research organization identifiers compels the necessity of addressing the affiliation disambiguation problem [6; 22]. Addressing this challenge is crucial for open science directives.

Authors that belong to the same research organization often mention their affiliations in different textual manners [5]. Reasons for affiliation ambiguity arise from various sources:

1. Misspelling, typographical errors, semantic expression of the institution and inconsistent formatting of institution names in bibliographic datasets play a major role in affiliation ambiguity [6; 13].
2. Researchers tend to affiliate themselves with multiple institutions for numerous reasons: strong incentives such as funding, ranking of the institution, casualization of the academic profession, and the decline in institutional support for academics [11]. Extracting multiple affiliations from text is difficult for traditional algorithms.
3. Institutions with identical names/abbreviations could also lead to ambiguity.

Alexandria3k is an open-source software library and command-line tool that allows performing systematic research on published literature through efficient querying of diverse open datasets (Crossref, ORCID, USPTO and more) [23]. This package efficiently mitigates volume, transparency, repeatability, and reproducibility issues. The research aims to evaluate and provide improvements to Alexandria3k in terms of identification and disambiguation of affiliations, and subsequently aid the researchers in their research. Alexandria3k utilizes existing meta-data and generates new tables with disambiguated author affiliation records.

The research question builds upon this premise for the Alexandria3k package, **How good is the existing author affiliation matching (based on naive maximal sub-string matching) in Alexandria3k, and how can it be improved?** The project scope includes establishing a baseline for the current “naive string-matching” strategy and proposing improvements through a novel approach. With the advancements made in Natural Language Processing (NLP) and Large Language Models (LLMs) such as GPT-4, we decided to integrate this technology to help with disambiguation. LLMs are effective in extracting contextual information and capturing semantic relationships; the precise technology to address our problem [10].

The research question could be divided into the following sub-questions:

1. **What is the baseline performance of the string matching algorithm in Alexandria3k when compared to the ground truth?**
   (a) How can the ground truth for author affiliations be created?
   (b) How can the currently implemented algorithm be compared against the ground truth?
   (c) What is the performance measure of this algorithm in terms of precision?

   The currently implemented author affiliation matching is based on a naive maximal sub-string matching algorithm, and optimized using the Aho-Corasick automaton. We need to understand how the authors are affiliated using this algorithm. We must translate this understanding into a systematic approach to create the ground truth. The process of author affiliation in Alexandria3k should be compared against the ground truth to establish a baseline for the system. The baseline performance can then be measured using precision. We need to identify the limitations and gaps in the dataset to grasp the context of the baseline performance.

2. **Can the use of a Large Language Model (GPT4) improve author affiliation linkage in Alexandria3k?**
   (a) What prompt can extract the affiliations from the text consistently?
   (b) How can the results of the LLM be verified?
   (c) How does the LLM perform in comparison to the existing string-matching algorithm?

   Using LLMs to decipher and extract information from large textual pieces is a sensible and proven method

[20]. We intend to leverage the characteristics of LLMs to improve affiliation identification and matching. Subsequently, we need to devise a prompt that would allow us to extract affiliation from the text in a consistent format. We should compare the results of our process to that of the baseline to verify the success of our proposed algorithm.

The rest of the paper is as follows. Section 2 summarizes previous work done in affiliation disambiguation. Section 3 describes the datasets used, the process of creation of ground truth, the establishment of the baseline and the approach of using LLMs to disambiguate affiliations. Section 4 describes the results of this research in sections of ground truth, baseline and results from our newly created process of affiliation disambiguation. Section 5 discusses the ethical aspects of our research. Section 6 compares our process to the previous works, highlights the limitations of our process and discusses the threats to validity. Section 7 summarizes the research and suggests directions for future improvements on this topic.

2 Related Work

This section aims to highlight the background and summarize the works in the field of affiliation disambiguation based on its approach. These works can be divided into 2 types of approaches: rule-based and using entity recognition. As none of the works reviewed utilize NLP to disambiguate affiliations, it is an unexplored avenue in this domain.

Background of Alexandria3k and Affiliation Disambiguation

“Open Reproducible Scientometric Research with Alexandria3k” by Spinellis Diomidis from 2023 is the parent research paper that the current research builds upon [23]. It discusses the importance of affiliation disambiguation for context and completeness in scientometric research. It details the currently implemented algorithm in Alexandria3k based on efficient string matching [1]. It also discusses the potential mismatches due to incompleteness and lack of overlap between bibliometric datasets.

The research paper “Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems” by Donner, Rimmert and van Eck from 2019 compares the performance of different institute name disambiguation systems using bibliometric indicator values as the metric [6]. The paper mentioned three institute name disambiguation systems; 1. Web of Science (WoS) normalized institute names and organization-enhanced system, 2. Scopus affiliation ID system and 3. Independent institution disambiguation system. The systems are assessed based on their precision, recall, indicator value and ability to identify publications to the correct research institute. The independent institution disambiguation system, which utilizes a combination of rule-based and machine-learning approaches, outperforms the other systems significantly. The limitation of this system is that it can’t be transferred to another country or region with a different institution structure and naming convention and requires significant resources which makes it inaccessible to researchers.

Rule-based approaches to Affiliation Disambiguation

The research paper “Efficient supervised and semi-supervised approaches for affiliations disambiguation” by Cuxac et al. from 2013 aims to tackle the problem of author disambiguation in bibliographic databases through the proposed supervised and semi-supervised approach based on the Naive Bayes (NB) classifier and overlapping clusters [5]. They discuss the field of named entity disambiguation and data standardization as the backbone of their research. Their research shows encouraging results for both supervised and semi-supervised approaches.

The research paper “Affiliation Disambiguation for Constructing Semantic Digital Libraries” by Jiang et al. from 2011 discusses affiliation disambiguation through a proposed clustering method based on normalized compression distance (NCD) [13]. The paper details about various data compressors that can be used, explain the agglomeration clustering algorithm using pseudocode and shows how the choice of data compressor can affect the results. The results show that NCD outperforms the traditional k-means method in terms of statistical metrics such as average precision, F-measure, entropy, and purity. NCD is resistant to noise and can measure the similarities between affiliations relatively precisely but produces a sub-optimal number of clusters due to the dynamic nature of affiliations. One of the main reasons why k-means loses its robustness is due to the use of a dictionary that addresses the issue of abbreviations in affiliations.

Entity Recognition as an approach to Affiliation Disambiguation

The research paper “Using Elasticsearch for entity recognition in affiliation disambiguation” by L’Hote and Jeangigard from 2021 used Elasticsearch, a modular search engine technology for searching, indexing and analysing large volumes of data, to perform automatic alignment of affiliations from publication meta-data [16]. It allows the user to customize the alignment criteria by choosing different strategies without re-initializing the indexes. Subsequently, it allows users to maintain control over the precision and recall.

The research paper “ELAD: An entity linking based affiliation disambiguation framework” by Shao et al. from 2020 proposes a knowledge-graph based entity linking learning framework that tackles the issue of ever-changing and ever-increasing data that hinders affiliation disambiguation [22]. The framework consists of pre-processing data, candidate generation, result selection and application. The candidate generation is based on the entity-linking algorithm supported by XLore which considers various aspects of the institution. The result selection tends to find the most likely institutional entity from the candidate set by considering the institutional context in the affiliation. The results show significant improvement in recall, precision and accuracy when compared to traditional approaches.

3 Methodology

This research revolves around establishing verified records of authors with their affiliations, determining the baseline performance of the matching algorithm currently implemented in Alexandria3k and improving author-affiliation linkage in Alexandria3k using LLMs. *RQ1* was labelled as quantitative research as it predominantly deals with numerical data and performs statistical evaluation. *RQ2* experiments with LLMs as a novel technology in the sphere of author affiliation disambiguation. *RQ2* uses methodologies such as prompt engineering, exploratory data analysis (EDA) and experimental research to achieve the above-mentioned goals. In general, the research involved a plethora of EDA and experimentation.

3.1 Datasets

Three open-source datasets were used during the research: Crossref[4], ORCID[24] and ROR[21]. 25% of "April 2022 Public Data File from Crossref" was randomly chosen as a sample from the Crossref dataset for the baseline. This constituted 39.1GB (6,702 zipped JSON files) of 167.99GB from the entire Crossref dataset. This sample size represents a 95% confidence interval of the dataset with a ±1.1% margin of error. 1% of the above-mentioned Crossref dataset was chosen to compare the newly created process to the existing algorithm in Alexandria3k. This sample size signifies a 90% confidence interval with a ±5% margin of error. The entirety of the "ORCID_2022_10_summaries.tar.gz" ORCID dataset from October 2022 was used. This dataset contained 29.8GB of compressed ORCID data. Finally, the "v1.17.1-2022-12-16-ror-data.zip" ROR dataset from December 2022 (version: v1.17.1) was used. This dataset contained 20.1MB of research organization meta-data in JSON format, which comprises 104,402 uniquely identifiable research organizations. These datasets were then populated into SQLite 3.34.0 databases for faster and easier queries.

3.2 Ground Truth Creation

Data exploration gave us a holistic overview regarding the structure of data, presence of missing/extreme values, and inter-relationships we were dealing with from the dataset[15]. Author affiliations across different datasets aren't conclusive, so creating a ground truth table was essential [7]. The assumption made whilst creating the ground truth was that "all" research organizations are identified and indexed by the Research Organization Registry (ROR)[21].

The ground truth for this research was created by exploiting the affiliations of authors available in the ORCID dataset. The ORCID dataset contains records of authors that are unambiguous and uniquely verified. It also contains information regarding their published works, qualifications, funding, employment, affiliations, etc. These affiliations/employments are research organizations that were cross-referenced across the Research Organization Registry using various organization identifiers to verify their validity. The ORCID dataset contained a variety of organization identifiers such as ROR, GRID, FunderRef, WikiData, ISNI and others. Appendix A.1 shows how we identified and stripped the identifiers using SQL. The above-mentioned identifiers were also supported in the ROR dataset. This process of creating and labelling the ground truth is visualized in Figure 1.

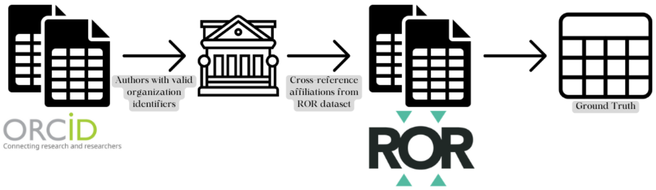
Authors with valid organization identifiers
ORCID
Connecting research and researchers
Cross-reference affiliations from ROR dataset
ROR
Ground Truth

Figure 1: Creation of ground truth

3.3 Baseline: Running author-affiliation process in Alexandria3k

The subsequent step was establishing a baseline by comparing the currently implemented algorithm (maximal sub-string matching) in Alexandria3k with the ground truth. The baseline performance is based on the number of matching records and the precision of the algorithm.

The process of author affiliation in Alexandria3k is a command through the command-line interface called "link-aa-base-ror". This process uses the affiliations in the Crossref database (table: author_affiliations) to match said affiliation to the organizations identified by ROR. 25% of the Crossref dataset was used to establish the baseline. The link is created using common sub-string matching on the name, aliases and acronyms of the research organization. This process is optimized using the Aho-Corasick automaton for fast and memory-efficient multi-pattern string search [18]. The process creates a table of author-organization pairs called "work_authors_rors" where the author is indexed from the work_authors table (Crossref database) and linked to their affiliated organization, which is indexed from the ROR database.

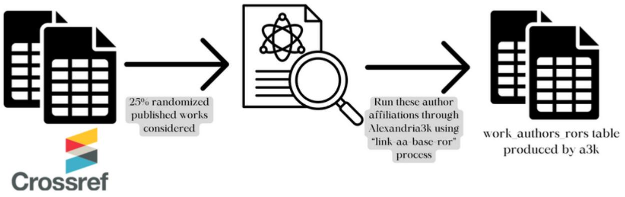
25% randomized published works considered
Crossref
Run these author affiliations through Alexandria3k using "link-aa-base-ror" process
work_authors_rors table produced by a3k

Figure 2: Visualization of author affiliation linkage process in Alexandria3k

The baseline is produced by overlaying the results from this process on the ground truth. We choose the author affiliations with valid ORCID from the "work_authors_rors" table. Using ORCID as the verified identifier, we can verify each author to their respective affiliation. This process is visualized in Figure 3. Using Persistent Identifiers (PIDs) enables us to trust the results. It is also the best way to methodically create and reproduce said results. We use this baseline to determine the algorithm's precision and matching rate.

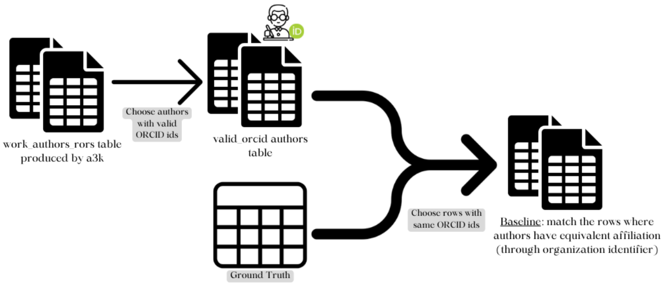
work_authors_rors table produced by a3k
Choose authors with valid ORCID ids
valid_orcid authors table
Ground Truth
Choose rows with same ORCID ids
Baseline: match the rows where authors have equivalent affiliation (through organization identifier)

Figure 3: Baseline of Alexandria3k

Another process called "link-aa-top-ror" is available in Alexandria3k which links authors with their parent organizations. We do not use this process for the research because we want to improve on the base author affiliation disambiguation.

3.4 Experimental Research: GPT-4

Through EDA, we noticed that the current algorithm fails to extract any relevant information when the affiliations texts contain additional affiliation jargon such as faculty, the role of the researcher, the address of the research organization, and multiple affiliations. Data extraction approaches that follow guidelines or use string matching are unfeasible for unstructured affiliations. We designed a process that uses LLMs to extract affiliations from the text. The use of LLMs addresses the limitations mentioned above. We chose to use the GPT-4 model from openAI as the most recent and state-of-the-art model [19].

We used iterative prompt engineering to write prompts that would return consistent and structured results for each affiliation. The initial results were too textual, so we had to include constraints in the prompt (`ONLY university/research organization and city`). The initial results were inconsistent, it would unpredictably switch the order of research organization and city. Hence we had to provide the format `(organization, city)` in which we wanted the response. We also had to instruct the model for edge cases and unreliable input. We used capitals for stronger emphasis on certain words. The final prompt used is mentioned below:

> From the given textual piece, identify ONLY the university/research organization and city mentioned. The response should be (organization, city). DO NOT produce any other textual information. Ignore all other information mentioned in the textual piece. If the organization or city is not recognized, then return (_,_). + "Textual affiliation"

We then shrink our search based on the city of the research organization. Once we have narrowed our search results to all the research organizations within the specified city, we use Levenshtein distance (calculates the string similarity) [2] to determine the best match. The process is visualized in Figure 4.

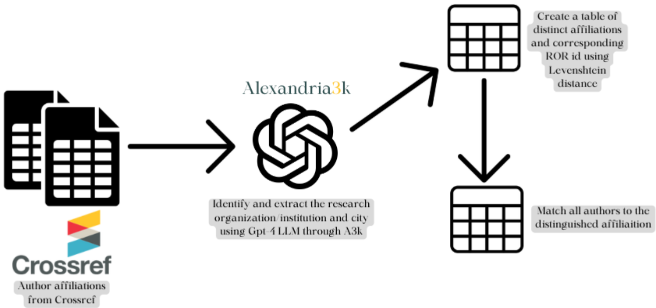
Alexandria3k
Crossref
Author affiliations from Crossref
Identify and extract the research organization/institution and city using Gpt-4 LLM through A3k
Create a table of distinct affiliations and corresponding ROR id using Levenshtein distance
Match all authors to the distinguished affiliaition

Figure 4: Author affiliation linkage using LLM

4 Results

This section contains the results of the experiments carried out throughout the research. The results are ordered based on the research questions. Section 4.1 gives us the number of author affiliations identified based on each organisation identifier. Section 4.2 indicates the performance of the existing algorithm in terms of raw numbers and statistical metrics. Section 4.3 shows how the LLM performs compared to the baseline model.

4.1 Ground Truth

As mentioned in Section 3.2, the ground truth was created using publicly available employment and education information from ORCID. Table 1 shows the authors that have some form of affiliations mentioned in their profile and the organizations that could be identified using an organization identifier (GRID, ROR, FunderID, Wikidata).

Entity	Records
Unique authors identified from ORCID	1,951,171
Authors with available affiliations	752,263
Authors with available affiliations and valid (not NULL) organization identifier	498,198
Affiliations identified by GRID	78,869
Affiliations identified by ROR	63,707
Affiliations identified by Funder_Id	33,255
Affiliations identified by Wikidata	1,617
Total #rows with valid organization identifier	177,448

Table 1: Number of records from ORCID dataset

Based on the records, 66.22% author affiliation pairs have an organization identifier. This represents the percentage of records whose organization identifier column is not `null`. The remaining records have textual affiliations but are not linked with any PID (Persistent Identifier). Furthermore, 23.58% of the records could be linked to a research organization through a valid recognized PID. The distribution is represented in Figure 5. Other PIDs mentioned in the ORCID database were unidentified and could not be used to match the research organizations. No organisations could be identified through the ISNI (International Standard Name Identifier) identifier. This is due to the unique 16-number formatting of ISNI that was not translated into the ORCID dataset.

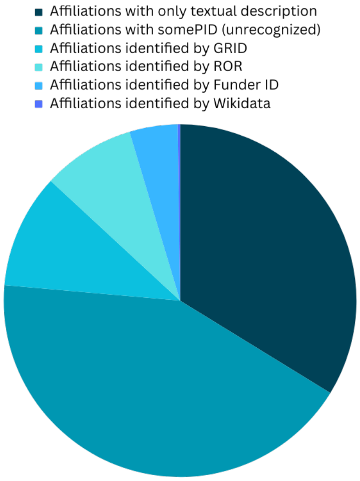

Figure 5: Distribution of author affiliation records in ORCID

4.2 Baseline

To create the baseline, we used 25% Crossref dataset and ROR dataset to create the “work_authors_rors” table which contains authors and their affiliations through Alexandria3k. Table 2 displays the number of affiliations identified by the Crossref dataset and the ones identified by Alexandria3k.

Entity	Records
Author affiliations available from Crossref	19,324,614
Records with valid ORCID	2,342,076
Records with distinct ORCID	980,312
Affiliations identified by A3k	7,290,985
Records in A3k with valid ORCID	929,729
Records in A3k with distinct ORCID	530,645

Table 2: Affiliations identified by Crossref and Alexandria3k

We observe that **37.72%** of the affiliations available through Crossref have been recognized by the common sub-string matching algorithm of A3k. This percentage shows the **matching rate** of the algorithm. This shows that $> 62\%$ of affiliations aren’t being identified due to the limitations of the currently implemented algorithm. Furthermore, $\sim 12\%$ of the author affiliation records in Crossref and identified by A3k have valid ORCID. Table 3 shows the number of records that were found in the ground truth, affiliations identified by A3k using ORCID (baseline) and successful recognition of affiliation in the baseline (matches).

Identifier	Ground Truth	Baseline	Matches
GRID	78,869	12,297	6,070
ROR_Id	63,707	6,703	2,333
Funder_Id	33,255	5,463	1,873
Wikidata	1,617	659	0

Table 3: Rows from ground truth, baseline and successful matches for every organization identifier used.

Precision is the number of correct predictions made:

$$Precision = TP/(TP + FP)$$

where TP is *True Positive* and FP is *False Positive*. For our case, precision is defined as the number of successfully identified research organizations from the total number of author affiliations identified from the ground truth. This gives a **precision** of **0.493**.

4.3 LLM Improvement

We use 1% of the Crossref dataset to compare the newly created process to the existing algorithm in Alexandria3k. A smaller subset is chosen due to the cost of running the process and proving the improvement of this approach. The process uses the openAI API to query results, which takes around $10 - 20$ seconds for each affiliation. The cost of running such a process on $\sim$ 19 million author affiliations records (25% of Crossref) would be impractical and expensive. However, the process of distinguishing affiliations is a one-time command and can be reused for other similar affiliations. The linkage of authors to these distinguished affiliations is completed in milliseconds. The details of this subset and the comparison of the Alexandria3k process to our process are described in Table 4.

Entity	Records
Author affiliation mentioned in Crossref	62,359
Records identified by A3k	22,905
Records identified by LLM process	50,675
Distinct affiliations mentioned in Crossref	25,214
Distinct affiliations identified by A3k	3,768
Distinct affiliations identified by LLM	20,599
Authors with multiple affiliations (Crossref)	6,835
Multiple affiliations identified by A3k	816
Multiple affiliations identified by LLM	3,973

Table 4: Raw number comparison between Alexandria3k process and LLM model

In line with the baseline, the **matching rate** of the process implemented in Alexandria3k is **36.73%**. In comparison, our process can identify and match **81.26%** of author affiliation. This matching rate is achieved due to the high number of distinct research organizations being extracted from affiliation texts. While the Alexandria3k process can identify only **14.94%** of the affiliations, due to the nature of the LLM information extraction process, we can achieve a substantial **81.69%** identification rate of the affiliations mentioned.

One of the important characteristics of this process is its ability to match an affiliation text containing multiple research organizations to each of them. This was a limitation observed in the Alexandria3k algorithm where **11.93%** of affiliation texts with multiple research organizations mentioned were identified and matched. Our algorithm has improved the identification rate almost 5-fold to **58.12%**.

Through the above-mentioned comparing metrics, we can establish that our algorithm results in significant improvements to author affiliations linkage in Alexandria3k.

5 Responsible Research

The research involves a lot of authorship information from various authors. The published datasets contain information that could be used to identify authors, their works, their location, and their affiliations. These seemingly disjoint pieces of information could collectively identify patterns in the person's work behaviour which is not mentioned in the metadata. Our algorithm processes affiliations in a modular manner to avoid any overlap of authorship information that could generate patterns of the authors' work.

Bias could emanate from LLMs whilst performing the process as LLMs are usually trained on uncurated heaps of Internet available data [9]. These biases could emanate from various sources: negative sentiments, linguistic associations, lack of recognition, etc [3; 12; 17]. Most LLMs are trained on English and other dominantly used languages, thus they tend to be proficient in these languages. This could induce *linguistic bias* in identifying research organizations from textual affiliation in a minority language or dialect. *Temporal bias* could seep into LLMs due to the unavailability or choosing to exclude training data from regions of conflict based on current events [8]. For example, a newly established research organization from a region of conflict may not be identified. We need to accept the inevitable presence of bias in LLMs. Writing unbiased prompts through prompt engineering could improve the consistency of results and inhibit the model from introducing inherent bias.

The process of integrating LLMs into Alexandria3k has been carefully documented and will be available in Github[1] for repeatability. The process, data formatting and prompts have been discussed in this paper for openness and reproducibility. Along with this, most of the SQL queries are also available in Appendix A. The datasets will be made available for future research and cross-referencing. The source code is openly accessible and licensed under GNU General Public License. These measures ensure that the research and source code align with the FAIR (Findable, Accessible, Interoperable, Reusable) principles.

6 Discussion

We aim to provide engaging insights into the process and results of our research. The nature of LLMs produces a few unexpected scenarios in disambiguating affiliations. We compare the approach of designing our process to previous works in this section. We also discuss the technical and theoretical limitations that affect the performance of our process and threats to the validity of this research.

Engaging Insights and Comparison to Previous Works

The existing author affiliation (based on naive maximal sub-string matching) has innate limitations and rigid matching criteria, reflected in its evaluation. The presence of affiliation-related jargon (department name, role of the researcher, address of the institution, stop words) from textual descriptions of affiliations hinders the matching accuracy of the existing algorithm. The use of LLMs in comparison displays a significant leap in performance in all relevant metrics. The substantial enhancements in identification rates and the ability to handle multiple affiliations affirm the improvement of our algorithm in linking author affiliations within the Alexandria3k framework and answers *RQ2*.

One of the key differences between our research and the previous works is the type of problem we are dealing with. Previous works dealt with multi-class classification, clustering and entity recognition problems. We dealt with a one-class classification problem [14], where we needed to assign an unknown element (affiliation) to the correct university/organization from the category of research organizations. For this reason, we are unable to calculate TN (True Negatives) and FN (False Negatives), and consequently, recall and F1 score for our algorithm.

Our research follows a similar strategy to affiliation disambiguation to that of ELAD Entity recognition research by Shao et al. [22] and to that of constructing semantic digital library research by Jiang et al. [13]. Shao et al. suggest a set of candidate institutions that could be matched to the provided affiliation and select the most likely candidate. Their candidate set is generated using knowledge graphs and related entities whilst ours is generated by filtering research institutes based on location. Similarly, Jiang et al. use NCD as a metric for clustering affiliations that are similar. Furthermore, ELAD's approach selects the most likely candidate based on the longest common subsequence whilst ours selects based on Levenshtein distance.

A paramount advantage of using LLMs is the quality of results despite the lack of context provided. The previous works require a broad spectrum of data to provide context for disambiguating affiliations. In comparison, LLMs perform this task with just the textual affiliation. However, affiliations with a severe lack of essential information regarding the research organization cannot be extracted correctly (for example, "Department of Psychiatry, Bolzano, Italy" doesn't explicitly mention any research organization that can be deduced by the LLM). The brittle and unpredictable nature of LLMs can often produce sub-par results for straightforward affiliations (for example, "Georgia Institute of Technology" is not recognized but "Georgia Institute of Technology, School of Civil and Environmental Engineering, Atlanta, Georgia, USA" is recognized).

A bottleneck of our process is the number of queries processed by openAI API and the significant amount of time required to run the entire process. This issue could be addressed by switching to an open-source, locally run LLM instead of using an API.

[1] https://github.com/dspinellis/alexandria3k

organization_name	organization_city	organization_identifier	start_year
University of Lisbon	Lisbon	https://ror.org/01c27hj86	2022
Universitat de Barcelona	Barcelona	$\prec null \succ$	2011
University of Bristol	Bristol	1980	2005
Universidade de Lisboa	Lisboa	37809	1984

Table 5: Affiliations of author A from ORCID

Limitations of Ground Truth

While curating the ground truth, about $2/5$ author affiliation records had valid organization identifiers. One organization identifier, RINGGOLD, was identified through manual inspection. However, this identifier could not be used to verify the organization because RINGGOLD does not have an openly accessible dataset. Hence, we can not cross-reference the linked research organization across ROR. The only way to verify the research organization with RINGGOLD is by manually querying through their website. This could imply that Alexandria3k and our algorithm correctly identified many more author affiliations but could not be matched to the ground truth due to the absence of these records.

Table 5 displays the affiliations of an author where one of the affiliations is identified by the ROR identifier whilst the others are not (table inserted directly from querying the ORCID database). The affiliation of this author mentioned in Crossref is displayed as:

```
Catalan Institute of Research and Advanced
Studies, 08010 Barcelona, Spain;

Department d'Història i Arqueologia (Grup
de Recerca SGR2014-00108), University of
Barcelona, 08010 Barcelona, Spain;
```

We notice that the affiliation mentioned in Crossref corresponds to when the author was affiliated with the University of Barcelona. However, since the author fails to include a valid organization identifier for “Universitat de Barcelona” in ORCID, we are unable to verify this systematically. There are several such examples when matching the results of Alexandria3k to the ground truth. This justifies the relatively poor performance of Alexandria3k.

Threats to Validity

An important threat to validity involves the accessibility of the datasets used in this research. The datasets used could be discontinued, removed or censored due to unforeseen circumstances in the future. This would make it impossible to repeat this research. A similar argument could be made about the continuation and usability of Alexandria3k and GPT-4. Verifying the results from GPT-4 can be difficult as the process is not run locally and a LLM can produce slightly different results in different iterations. Maturation and advancements in LLMs could pose a threat to validity as well. The research performance could become obsolete if a LLM which is more powerful and accurate than the current model is introduced. The methodology of our research would remain valid and relevant, but the results could improve drastically. Furthermore, as discussed before, different iterations of running the process could produce different accuracy of results. Hence, the identifying and matching rates need to be considered with a margin of error.

7 Conclusion and Future Work

This research aimed at contextualizing and measuring the performance of the existing author affiliation algorithm in Alexandria3k and using LLMs to improve the disambiguation of affiliations for enhanced results. Forming a baseline using the ground truth highlights 36.73% matching rate of the algorithm with a precision of 0.493. We proposed an algorithm to leverage the entity extraction capabilities of LLMs (GPT-4) to improve author affiliation linkage in Alexandria3k. Our process filters and extracts the affiliations of authors from textual descriptions and matches them to their corresponding research organization/institute. Our process performs significantly better in identifying organizations from textual descriptions and linking authors to multiple affiliations. Our algorithm achieves a matching rate of 86.26% and an identification rate of 81.69%. We also discuss the similarity of our methodology to that of Shao et al. (ELAD entity recognition) and Jiang et al. (constructing semantic digital libraries). A significant bottleneck in the performance of our algorithm lies in using the OpenAI API to distinguish affiliations. Our research aims to establish the success of using LLMs and open the door for further research in the domain of affiliation disambiguation using LLMs.

There are a plethora of paths to explore in the domain of affiliation disambiguation and Alexandria3k.

- The integration of other organization identifiers (RINGGOLD) can produce an extensive ground truth. Introducing flexible and liberal criteria for the inclusion of research organizations without an organization identifier can aid the creation of an expanded and complete ground truth.
- The implementation of approaches such as NCD, ELAD, and Elasticsearch from Section 2 in Alexandria3k could be useful for system users. Alexandria3k could serve as a common environment for comparing different approaches to disambiguate affiliations.
- Implementing the process mentioned in this research using open-source language models such as Phi-2 (from Microsoft), Mistral (from MistralAI) and LLama (from Meta) could enable the users to run the process locally. A broader context could be provided to link authors to affiliations directly.

References

[1] Aho, A. V., and Corasick, M. J. Efficient string matching: An aid to bibliographic search. *Commun. ACM 18*, 6 (jun 1975), 333–340.

[2] Bachmann, M. python-levenshtein 0.23.0: Python extension for computing string edit distances and similarities., 2023. License: GNU General Public License v2 or later (GPLv2+) (GPL).

[3] Blodgett, S. L., and O'Conner, B. Racial disparity in natural language processing: A case study of social media african-american english. *ArXiv* (2017).

[4] Crossref. April 2022 public data file from crossref.

[5] Cuxac, P., Lamirel, J.-C., and Bonvallot, V. Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics 97* (2013), 47–58.

[6] Donner, P., Rimmert, C., and van Eck, N. J. Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies 1*, 1 (02 2020), 150–170.

[7] Fang, J., Tao, X., Tang, Z., Qiu, R., and Liu, Y. Dataset, ground-truth and performance metrics for table detection evaluation. In *2012 10th IAPR International Workshop on Document Analysis Systems* (2012), pp. 445–449.

[8] Ferrara, E. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv* (2023).

[9] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *ArXiv* (2023).

[10] Gupta, G., Rastegarpanah, B., Iyer, A., Rubin, J., and Kenthapadi, K. Measuring distributional shifts in text: The advantage of language model-based embeddings. *ArXiv* (2023).

[11] Hottenrott, H., Rose, M., and Lawson, C. The rise of multiple institutional affiliations in academia. *Journal of the association for information science and technology 72* (2020).

[12] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Association for Computational Linguistics, pp. 5491–5501.

[13] Jiang, Y., Zheng, H.-T., Wang, X., Lu, B., and Wu, K. Affiliation disambiguation for constructing semantic digital libraries. *Journal of the American Society for Information Science and Technology 62* (2011), 1029–1044.

[14] Khan, S. S., and Madden, M. G. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review 29*, 3 (2014), 345–374.

[15] Kotu, V., and Deshpande, B. Chapter 3 - data exploration. In *Predictive Analytics and Data Mining*, V. Kotu and B. Deshpande, Eds. Morgan Kaufmann, Boston, 2015, pp. 37–61.

[16] L'Hôte, A., and Jeangirard, E. Using elasticsearch for entity recognition in affiliation disambiguation. *CoRR abs/2110.01958* (2021).

[17] Mozafari, M., Farahbakhsh, R., and Crespi, N. Hate speech detection and racial bias mitigation in social media based on bert model. *ArXiv* (2020).

[18] Muła, W., Ombredanne, P., Grigorev, A., Woakes, D., and Betts, E. pyahocorasick. https://github.com/WojciechMula/pyahocorasick/, 2022.

[19] OpenAI. openai 1.6.1, 2023. License: Apache Software License.

[20] Patiny, L., and Godin, G. Automatic extraction of fair data from publications using llm. *ChemRxiv* (2023).

[21] ROR. Research organization registry data (v1.17.1) [data set]. *Zenodo* (2022).

[22] Shao, Z., Cao, X., Yuan, S., and Wang, Y. Elad: An entity linking based affiliation disambiguation framework. *IEEE Access 8* (2020), 70519–70526.

[23] Spinellis, D. Open reproducible scientometric research with Alexandria3k. *PLoS ONE 18*, 11 (Nov. 2023), e0294946.

[24] Westwood, G. Orcid public data file 2022. *ORCID* (2022).

[25] Xin, G., and Blackmore, K. L. Recent trends in academic journal growth. *Scientometrics 108* (2016), 693–716.

A SQL Queries

The research involved a lot of data pre-processing, manipulation and management. Most of the publication, author and affiliation meta-data were accessed, filtered, and used through databases. For this purpose, we used SQL queries instead of Python scripts for simplistic instant results for various experiments. The following appendix contains the queries used in this research.

A.1 Identifying Organization Identifier

The first step of creating the ground truth was preparing the data. The ORCID dataset had a mixture of organization identifiers without any naming convention or identification record. We had to comb through the entire dataset to identify which organization identifiers were valid and which category they belonged to (ROR, GRID, Funder Id, WikiData and ISNI). Provided below are the SQL queries for this process:

```
-- Create a new table from person_employments
-- with filtered and stripped values (Funder Id)
CREATE TABLE person_employments_filtered_2 AS
    SELECT *,
        REPLACE(
            REPLACE(organization_identifier,
            'http://dx.doi.org/10.13039/', ''),
            '"', '')
            AS stripped_organization_identifier
    FROM person_employments
    WHERE
        organization_identifier LIKE
        'http://dx.doi.org/10.13039/%';
```

Similar to this query, we create tables for ROR, Wikidata, and GRID identifiers. We strip $https://ror.org/$ for ROR identifiers, $grid.$ for GRID identifiers and add Q for wikidata identifiers. For ISNI identifiers, we need to add 0s in front of the number till it reaches a length of 16.

```
-- Create a new table with modified
-- organization_identifier
CREATE TABLE person_employments_filtered_4 AS
SELECT *,
    CASE
        -- Do not modify for specified patterns
        WHEN organization_identifier LIKE
        'http://dx.doi.org/10.13039/%'
        OR organization_identifier LIKE 'grid.%'
        OR organization_identifier LIKE
        'https://ror.org/%'
        THEN organization_identifier
        ELSE
            SUBSTR('0000000000000000' ||
            organization_identifier, -16)
        END AS modified_organization_identifier
FROM person_employments
WHERE
    organization_identifier NOT LIKE
    'http://dx.doi.org/10.13039/%'
    AND
    organization_identifier NOT LIKE 'grid.%'
    AND
    organization_identifier NOT LIKE
    'https://ror.org/%' AND
     organization_identifier is NOT NULL;
```

Once we have extracted all the organization identifiers, we cross-reference the research organizations through the ROR database. This allows us to map every valid research organization that we identified to an ROR_Id in our database. We perform this query for every organization identifier.

```
-- Create table where organization identifier is
-- matched to their ror_id
CREATE TABLE person_grid_organization_mapping AS
    SELECT
        pe.person_id,
        pe.organization_name AS
        person_organization,
        ro.combined_grid AS
        research_organization_grid_id,
        ro.ror_id
    FROM person_employments_filtered_5 pe
    JOIN ror_grid ro
    ON
        ro.combined_grid =
        pe.organization_identifier;
```

A.2 Comparing A3k to Ground Truth

To obtain the baseline, we compare the results of A3k to the ground truth. Authors with valid ORDID were chosen from the results of A3k. The following query filters authors with valid ORCID:

```
-- This query is for finding the authors with
-- valid orcid ids and storing them in a table
CREATE TABLE valid_orcid_authors AS
    SELECT *
    FROM work_authors_rors
    INNER JOIN
        work_authors
    ON
        work_authors_rors.work_author_id =
        work_authors.id
    WHERE work_authors.orcid is NOT NULL;
```

We use the *valid_orcid_authors* table to compare the author affiliation pairs to the ground truth. The table matches the authors based on their ORCID and includes records that have their affiliations correctly matched. The following query reflects this action:

```
-- Sql query for creating baseline:
CREATE TABLE baseline_grid_id AS
    SELECT
        va.ror_id,
        va.orcid,
        pg.person_organization AS organization
    FROM valid_orcid_authors va
    JOIN
    person_grid_id_organization_mapping pg
    ON
        va.orcid = pg.orcid AND va.ror_id =
        pg.ror_id;
```