AUTHOR:
DIBYENDU GUPTA
D.T.GUPTA@STUDENT.TUDELFT.NL

TUDelft

USE OF LLMS TO IMPROVE AFFILIATION DISAMBIGUATION IN ALEXANDRIA3K

SUPERVISORS:
DIOMIDIS SPINELLIS
GEORGIOS GOUSIOS
KOEN LANGENDOEN

1. INTRODUCTION

- **Challenge:** Affiliation disambiguation from bibliographic databases
- Reasons for this problem [1]:
  - Heterogeneity of datasets
  - Outdated storage methods
  - Emergence of new research organizations
  - No globally accepted organization identifier
- Reasons for disambiguity of affiliations:
  - Misspelling, typos, semantic expression, inconsistent formatting
  - Multiple affiliations for an author
  - Identical names/abbreviations of organizations
- Alexandria3k (A3k) - open-source library for performing systematic research on published datasets

2. RESEARCH QUESTIONS

How good is the existing author affiliation matching, (based on naive maximal sub-string matching) in A3k, and how can it be improved?

RQ1: What is the baseline performance of the string the matching algorithm in Alexandria3k when compared to the ground truth?

RQ2: Can the use of a Large Language Model (GPT4) improve author affiliation linkage in Alexandria3k?

3. METHODOLOGY

3.1 Ground Truth

- Independent of other datasets
- Organization Identifiers used:
  - Funder Id
  - GRID
  - ISNI
  - ROR Id
  - Wikidata

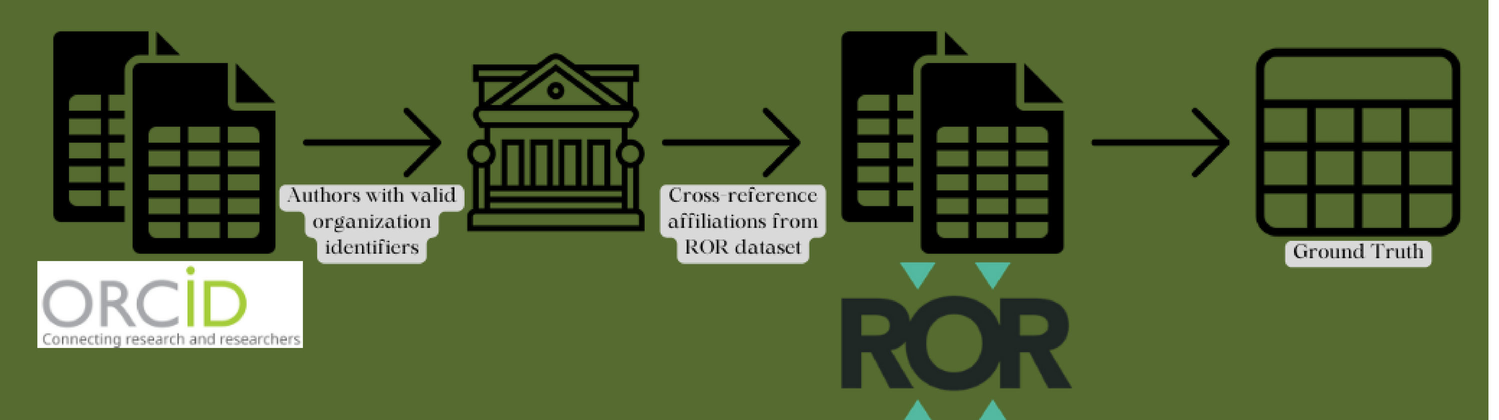

Figure 1: Process of creating the ground truth

3.2 Baseline of A3k

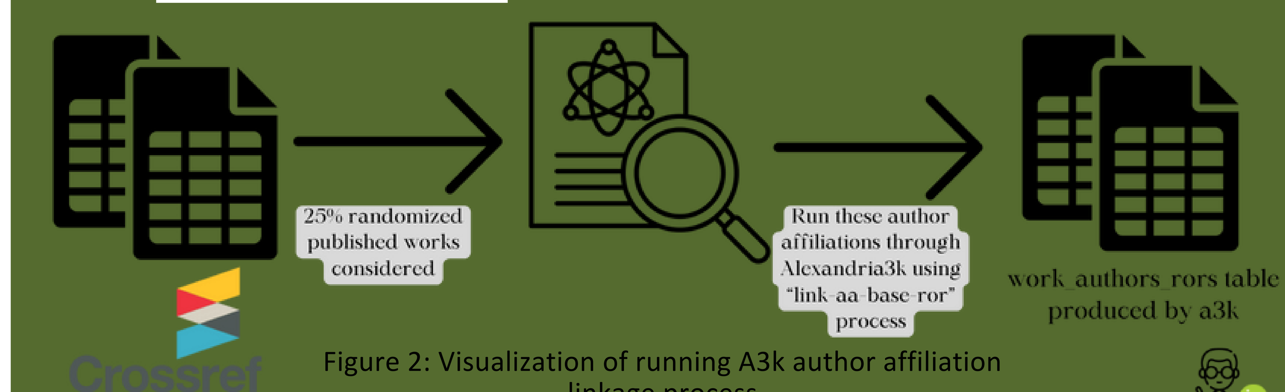

Figure 2: Visualization of running A3k author affiliation linkage process

Existing algorithm visualized:

- Using 25% of crossref author affiliations for the baseline
- Common sub-string matching on the name, aliases and acronyms of research organizations
- Optimized using Aho-Corasick automaton
- work_authors_rors table contains records of author and their respective affiliation

Baseline creation visualized:

- Comparing author affiliation pairs from A3k process to the ground truth
- Choose authors with valid ORCID for representative comparison
- Assumption: all organizations are identified and indexed by Research Organization Registry (ROR)

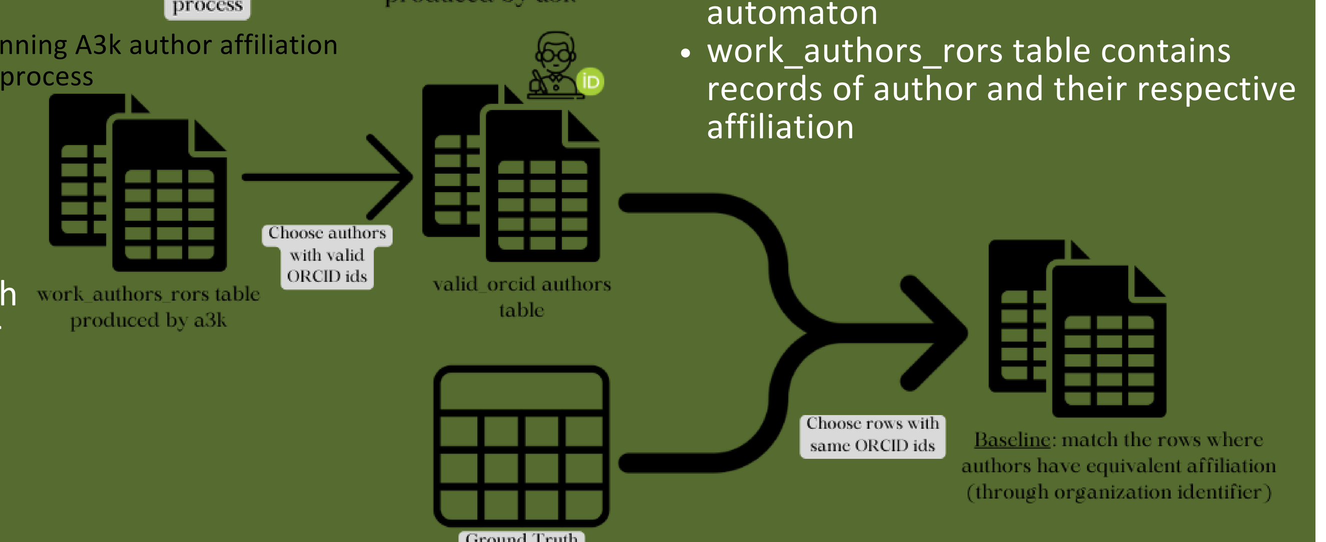

Figure 3: Visualization of comparing A3k process to ground truth

3.3 Using LLM (GPT-4) for affiliation disambiguation

- Use GPT-4 to extract affiliation from textual description (through OpenAI API) using prompt
- Process and format the response from GPT-4
- Use Levenshtein distance (calculates the string similarity) to find the best matching organization

Prompt used:

> 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"

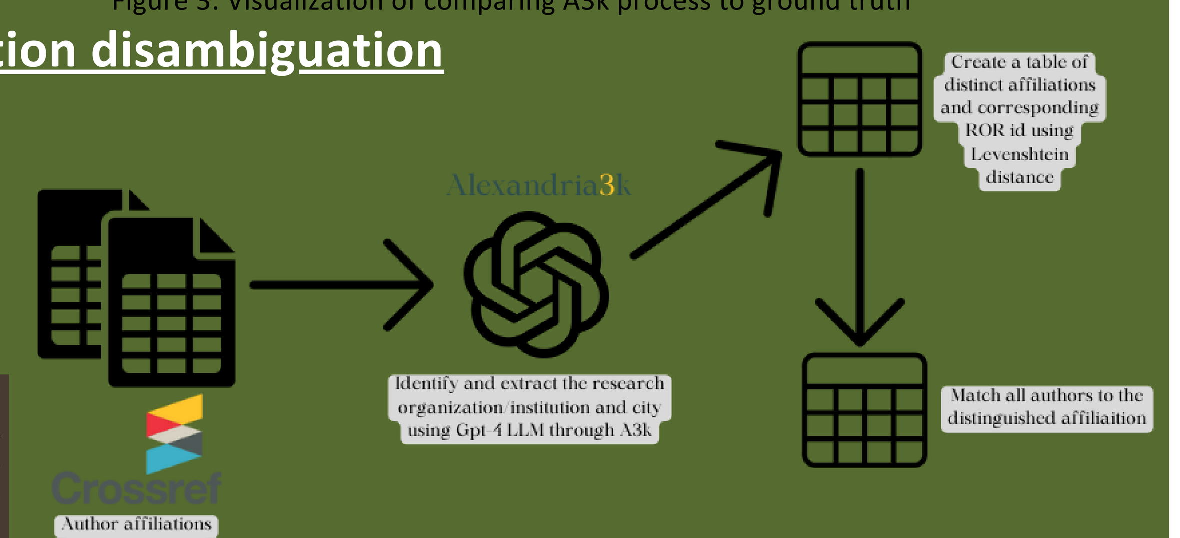

Figure 4: Process of author affiliation linkage using LLM

4. RESULTS

4.1 Ground Truth

- 66.22% of affiliations had a valid (not null) organization identifier column
- 33.78% of affiliations had only textual descriptions
- Organization identifiers identified: ROR, GRID, Wikidata, Funder Id

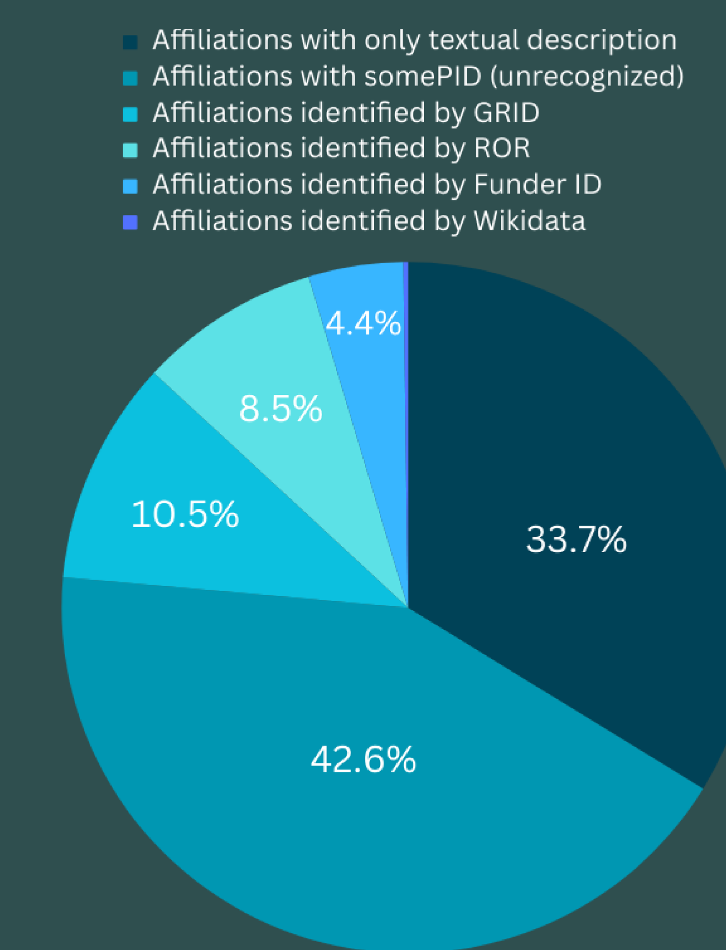

Figure 4: Ground Truth Distribution

4.2 Baseline Evaluation

- Performed on 25% of Crossref dataset
- Matching rate = 37.72%
- Precision = 0.493

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

Figure 5: Baseline-Comparing A3k process to ground truth

4.3 LLM Improvement

- Performed on 1% of the Crossref dataset
- Sample size signifies a 90% confidence interval with a ±5% margin of error
- Matching rate refers to the number of records (author-affiliation pair) identified
- Matching rate of A3k = 36.73%
- Matching rate of LLM = **81.26%**
- Affiliation identification rate of A3k = 14.94%
- Affiliation identification rate of LLM = **81.69%**
- Multiple affiliation identification in A3k = 11.93%
- Multiple affiliation identification in LLM = **58.12%**

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	50,599
Authors with multiple affiliations (Crossref)	6,835
Multiple affiliations identified by A3k	816
Multiple affiliations identified by LLM	3,973

Figure 6: Comparison between A3k and LLM

5. DISCUSSION

- Issue discussed in previous works: Multi-class classification problem Issue dealt in our research: one-class classification problem [2]
- Approach for affiliation disambiguation is similar to:
  - Shao et al.: creating candidate set and result selection using longest common subsequence[3]
  - Jiang et al.: normalized compressed distance (NCD) used to cluster affiliations [4]
- Brittle and unpredictable nature of LLMs:
  - Unable to recognize affiliation due to lack of essential affiliation information (ex: Department of Psychiatry, Bolzano, Italy)
  - Sub-par results for straightforward cases (ex: "Georgia Institute of Technology, School of Civil and Environmental Engineering, Atlanta, Georgia, USA" is recognized but "Georgia Institute of Technology" is not)
- Limitations:
  - Using OpenAI API affects performance in disambiguating affiliations
  - RINGGOLD organization identifier can not be identified in ORCID and referred to in the ground truth. No openly available datasets
  - In the example provided, ORCID is missing an organization identifier for "Universitat de Barcelona". This means that there is no record of the author being affiliated with Barcelona in the ground truth. So even when our process can disambiguate the textual affiliation from Crossref, we are unable to verify it.

Textual affiliation given in Crossref:

1. Catalan Institute of Research and Advanced Studies, 08010 Barcelona, Spain
2. Department d'Hist`oria i Arqueologia (Grup de Recerca SGR2014-00108), University of Barcelona, 08010 Barcelona, Spain

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

Figure 7: Affiliations of author A from ORCID

6. CONCLUSION & FUTURE WORK

- We have successfully improved author affiliation linkage in Alexandria3k using LLMs
- Our algorithm works exceptionally well in identifying distinct affiliations
- Directions for future work:
  - Integration of other organization identifiers such as RINGGOLD to expand the ground truth
  - Implement other approaches to affiliation disambiguation in Alexandria3k to compare the performance of different approaches, would make Alexandria3k a testing environment
  - Implement the LLM using open-source locally run models such as Phi-2, Mistral and LLama. It would mitigate a few of the limitations mentioned above.

7. REFERENCES

[1] 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.
[2] 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.
[3] SHAO, Z., CAO, X., YUAN, S., AND WANG, Y. Elad: An entity linking based affiliation disambiguation framework. IEEE Access 8 (2020), 70519–70526.
[4] 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.