Contributions to Alexandria3k

Author Name Disambiguation Using Large Language Models

Author:
Jelle van Lieshout

Supervisors:
Diomidis Spinellis
Georgios Gousios

References

[1] Z. Boukhers, N.B. and Asundi, Deep author name disambiguation using DBLP data. International Journal on Digital Libraries (2023).
[2] Diomidi 2023. *Alexandria3k documentation*. https://dspinellis.github.io/alexandria3k/.
[3] L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner, ORCID: A system to uniquely identify researchers. Learned Publishing 25, 4 (10 2012), 259–264.
[4] Big Science Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili, Daniel Hesslow, Roman Castagne, Alexandra Sasha Luccioni, et al. BLOOM: A 176B-Parameter Open-Access Multi- lingual Language Model. 11 2022.
[5] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science.

Conclusions & Limitations

- **LLMs are effective for AND**, outperforming state-of-the-art approaches.
- Mistral, a relatively small model, shows **superior precision, recall and F-score.**
- Performance declines when key **informative attributes** are missing.

Future Work

- Future improvements could come from more effective candidate selection, as current method misses potential matches (e.g.,incomplete first names).
- Adjustments in LLM configuration parameters, including temperature and system prompts, could further enhance performance.
- Using LLMs specifically trained on scientific publication datasets, like BLOOM [4] or Galactica [5], may yield more accurate matches due to their specialised knowledge.

Research Question & Hypothesis

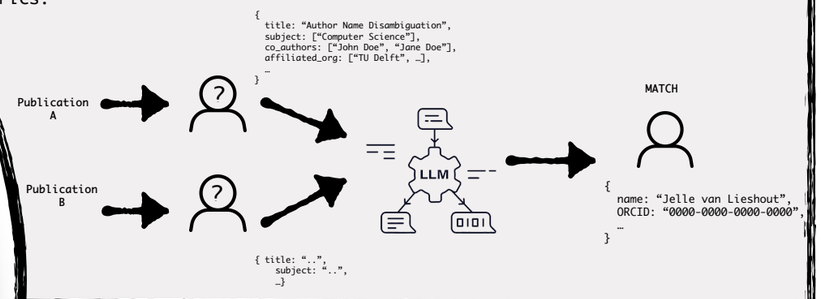
What is the comparative accuracy of large language models, such as llama2, in disambiguating author names within the CrossRef dataset, measured against the current state-of-the-art approach in terms of precision, recall, and F1 score?

Subquestions relate to generalisation, implementation possibilities and computational performance.

Hypothesis: large language models (LLM's) can predict author ambiguity more accurately than the current state-of-art approach.

Approach

- **Preparing and pre-processing:** Extracting key information from publication records (author names, co-authors, journal, organisation, title, subjects).
- **Prompting and interpreting output:** Presenting the model with two publication records and determining a match or non-match.
- **Logging and saving** the results from the model.
- **Repeating the process** for all relevant publication record combinations.
- **Validating matches** using a ground truth source and calculating performance metrics.



Background

Alexandria3k (a3k) is a software system providing local relational query access to diverse publication open data sets [2].

Author Name Disambiguation (AND) refers to the process of establishing whether two authors with the same first and last name, are also the same real-world person.

CrossRef is an openly accessible publication database containing 60+ million journal studies.

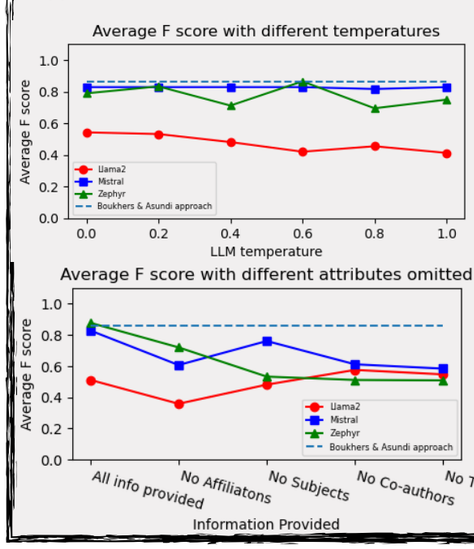
Example: in the *DBLP* dataset, there are 37,409 publications referring to authors with atomic name variate 'Y Zhang'. There are only 2601 unique authors with 'Y Zhang' as an atomic name variate [1].

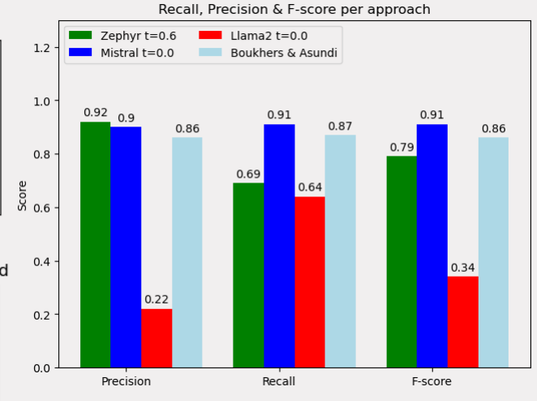
State-of-the-art Method [1]

Extract atomic name variates from publication records → Group records by atomic name variate → Vectorize author, co-author and publication info → Classify to author using DNN model per name variate

Results

Experiments using the approach proposed by Boukhers et al [1] against the novel LLM approach, when implemented in a3k and tested on a 10% random sample of an ORCID-labelled [3] CrossRef dataset for atomic name variate 'Y Zhang'. Experiments are run to find optimal configuration as well as comparing against Boukhers & Asundi approach.





As seen in the bar chart, *Mistral temp=0.0* provides a very suitable candidate to disambiguate author names, outperforming Boukhers and Asundi on every metric.

TU Delft