Independent Study Report: Name Disambiguation in the Human Reference Atlas Scientific Literature (HRA Lit) Knowledge Graph

By: Aishwarya Mocherla
Professor: Michael Ginda
Summer - Second Six-Weeks : 2024

Course Outline

- **Credit Hours:** 2 credit hours
- **Number of Hours Expected to Work Each Week:** 8 hours
- **Meetings:** Weekly meetings to discuss progress
- **Contact Information:**
    - **Aishwarya Mocherla:** aimoch@iu.edu
    - **Michael Ginda (Instructor):** mginda@indiana.edu
- **Onboarding Document** : MocherlaAishwarya-Onboarding

Learning Objectives

The primary objective of this independent study is to address one of the major challenges in developing the Human Reference Atlas Scientific Literature (HRA Lit) knowledge graph: identifying unique sets of persons and organizations associated with research publications. Accurately identifying individuals and organizations enhances the fidelity of research literature searches, aiding in expert identification for collaborations, reviews, consultations, advice, and research evaluation activities.

Name disambiguation will be the core focus, with an emphasis on implementing accurate and reliable data processing workflows to uniquely identify authors and their affiliated organizations using advanced machine learning methods from relevant literature. The project will utilize citation data from the PubMed database maintained by the Cyberinfrastructure for Network Science Center.

The final project outcome will include:

- Data set of authors with unique IDs
- Github Repository with all the resources required to execute the project
- Documentation of the methodology and results for reproducibility and future research

Weekly Breakdown

- **Week 1 (17th June - 23rd June):** Literature Review
  - Onboarding and introductory presentation
  - Conduct an extensive literature review to understand current name disambiguation research, methodologies, findings, and gaps.
- **Week 2 (24th June - 30th June):** Methodology Design
  - Refine the research question and methodology based on the literature review. Finalize the project approach.
- **Weeks 3 & 4 (1st July - 14th July):** Algorithm Development
  - Start coding and implementing machine learning algorithms for name disambiguation. Document initial findings.
  - Intermediate Project Results Presentation
- **Week 5 (15th July - 21st July):** Testing and Refinement
  - Test various name disambiguation methods against each other. Analyze results and draft documentation on methodologies and findings.
  - Final Presentation I
- **Week 6 (22nd July - 26th July):** Final Analysis and Documentation
  - Complete the final analysis and consolidate all findings into a comprehensive report. Prepare for the presentation and discuss potential improvements.
  - Final Presentation II

Data Sources

- **GitHub Repository:** HRA Lit Data
- **Data Modeling:** SQL Data Modeling HRA Lit

Prior Work

- Prior work by Kiki
- QSS paper, https://doi.org/10.1162/qss_a_00299
- Nature paper (accepted) preprint at https://www.biorxiv.org/content/10.1101/2023.10.21.563417v1
- Building a PubMed knowledge graph. Jian Xu, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F. Rousseau, Xin Li, Weijia Xu, Vetle I. Torvik, Yi Bu, Chongyan Chen, Islam Akef Ebeid, Daifeng Li & Ying Ding. Scientific Data volume 7, Article number: 205 (2020) https://www.nature.com/articles/s41597-020-0543-2

- Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023 Feb 2;10(1):67. doi: 10.1038/s41597-023-01960-3. PMID: 36732524; PMCID: PMC9893183. https://pubmed.ncbi.nlm.nih.gov/36732524/
- Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, *47*(2), 227-254. https://doi.org/10.1177/0165551519888605
- Bib2Auth: Deep Learning Approach for Author Disambiguation using Bibliographic Data https://ar5iv.labs.arxiv.org/html/2107.04382
- Pairwise Learning for Name Disambiguation in Large-Scale Heterogeneous Academic Networks. https://ar5iv.labs.arxiv.org/html/2008.13099v2
- Citation-based bootstrapping for large-scale author disambiguation https://onlinelibrary.wiley.com/doi/full/10.1002/asi.22621
- A Graph-Based Author Name Disambiguation Method and Analysis via Information Theory. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7516896/
- Random Forests. https://link.springer.com/article/10.1023/A:1010933404324

Methodology

Data Merging and Preprocessing

Loading Data: Loading data involves importing the datasets from CSV files into Pandas DataFrames, which allows for efficient manipulation and analysis of the data. Data from three CSV files (`id.csv`, `name.csv`, `affiliation.csv`) was loaded into Pandas DataFrames. This step brought all necessary information into a workable format.

Standardizing and Filling Missing Data: Data standardization ensures consistency across the dataset. Author names were standardized to lowercase to ensure consistency, and missing ORCID entries were filled with a placeholder (`'unknown'`). This preprocessing step handled variations and missing values in the dataset effectively.

Data Merging: Data merging combines multiple DataFrames based on common identifiers, integrating different aspects of the data into a single DataFrame. The DataFrames were merged on common identifiers (`auth_id`, `pmid`). This step integrated various aspects of the authors' information into a comprehensive dataset for further processing.

Unique Identifier Creation: Creating unique identifiers ensures that each author entry is distinct, even if names are similar or identical. A unique identifier for each author entry was created by combining `pmid` and `auth_id`. To manage length and ensure uniqueness, a hash function was used to shorten these identifiers.

Handling Missing Values: Handling missing values is crucial for maintaining data integrity and ensuring the model can process the dataset without errors. Remaining missing values were filled with `'unknown'` to ensure the dataset was complete and ready for model training.

Removing Duplicates: Removing duplicates prevents the same author from being represented multiple times, which could skew results. Duplicates were removed based on the shortened unique ID to ensure the integrity of the unique identifiers.

Feature Engineering and Model Training

Feature Engineering: Feature engineering involves creating new features or modifying existing ones to improve the performance of machine learning models.LabelEncoder was used to transform categorical features (`auth_id`, `affiliation`, `fore_name`, `last_name`) into numerical values, which was necessary for the machine learning model to process the categorical data effectively.

Feature Selection: Feature selection is the process of identifying the most relevant features for the model, which helps in improving model accuracy and reducing complexity. The selected features for the model were `auth_id`, `affiliation`, `fore_name`, `last_name`. These features were chosen based on their relevance to the author disambiguation task.

Splitting Data: Splitting data into training and test sets is essential for evaluating the model's performance. The data was split into training and test sets using a 70-30 split.This split allowed the model to be trained and evaluated on unseen data.

Model Selection: Model selection involves choosing the appropriate machine learning algorithm based on the problem at hand.The Random Forest classifier was chosen for its robustness to overfitting, ability to handle a large number of input variables, and its ensemble learning nature. This model combines multiple decision trees to improve accuracy and robustness, making it suitable for the complex dataset..

Model Training: Model training is the process of feeding the training data into the machine learning algorithm to learn patterns and relationships. The Random Forest model was trained with 100 estimators. This training involved fitting the model to the training data to learn patterns and relationships within the features.

Evaluation: Model evaluation assesses the performance of the machine learning model using various metrics. The model was evaluated using accuracy score and classification report. These metrics provided insights into the model's performance, highlighting areas where the model excelled and where improvements were needed.

Saving Results: Saving the results involves storing the model predictions in a format that can be used for further analysis. The predicted unique identifiers for the authors were saved to a CSV file, including relevant author details. This ensured that the processed data was documented and ready for further analysis or use in the HRA Lit knowledge graph.

Challenges and Solutions

Handling Large Data: Handling large datasets can be challenging due to memory constraints and processing time. Work was limited to 100 entries to establish a base model due to the large dataset size. This approach allowed the methodology to be tested and refined before scaling up using chunk processing.

Machine Learning Model Selection: Choosing the right machine learning model is critical for achieving high accuracy. A custom Random Forest model was chosen due to its ability to handle diverse and complex features relevant to author disambiguation. This model's ensemble nature and robustness to overfitting were crucial for the project's success.

Performance Metrics Warnings: Warnings about precision and recall being ill-defined for certain labels were encountered. These warnings often occur when some classes have no true or predicted samples. This was addressed by focusing on improving data preprocessing and feature selection to ensure balanced classes and accurate model predictions.

Integration of Additional Methods: Exploration of using Large Language Models (LLMs) for data cleaning was conducted to enhance input data quality. Techniques like tokenization, noise removal, normalization, and lemmatization were considered to improve the overall performance of the model.

ORCID as the gold standard: Initially, ORCID identifiers were considered for better accuracy. Testing with data that included ORCID IDs showed better accuracy, but due to the limited availability (only about 30% of the authors had ORCID IDs), it was decided to proceed without relying on them extensively. This decision ensured the approach could be generalized to the entire dataset.

Future Work

Scaling Up: Implementing chunk processing to handle the entire dataset, testing model scalability and efficiency to ensure robustness.

Advanced Model Development: Exploring more advanced machine learning models such as Graph Neural Networks (GNNs) and Heterogeneous Information Networks (HINs) will be explored to improve disambiguation accuracy.

LLM Integration: Data cleaning will be refined using LLMs to enhance input quality and model performance. LLMs will be applied for advanced named entity recognition and resolution tasks.

Validation with External Standards: Authors with ORCID IDs and institutions with ROR IDs will be used as a gold standard for validating and optimizing the approach, as suggested by feedback received.