Navigating the Challenges of Deduplicating Actors in Historical Letter Exchanges

Senka Drobac[1], Petri Leskinen[1,2]and Muhammad Faiz Wahjoe[1]
[1]Aalto University, Espoo, Finland
[2]University of Helsinki, Finland

senka.drobac@aalto.fi
petri.leskinen@aalto.fi
muhammad.wahjoe@aalto.fi

Abstract: This paper outlines ongoing research to standardize the metadata of letters from Finnish cultural heritage organizations to build a comprehensive archive for studying 19th-century epistolary culture in the Grand Duchy of Finland (1808/09-1917), as a part of the Constellations of Correspondence (CoCo) project. The challenge of deduplicating and reconciling actors and their associated information is an important problem faced in historical knowledge databases. The actors, e.g., senders and recipients of letters, can be individual people, couples, or families as well as organizations like institutions or companies. The deduplication described in this article has its focus on the individual people. The solutions we propose in this paper are significant for the development of the efficient and accurate approaches in this field.

Keywords: Actor deduplication, Historical correspondence, Epistolary data, Semantic web, Knowledge graphs

1. Introduction

The initial phase of the Constellations of Correspondence (CoCo) consortium (Tuominen et al, 2022) has involved collecting metadata on 19th-century correspondence from different institutions across Finland. This data was generated in various formats and is currently undergoing standardization, as outlined in Drobac et al (2023), to conform to a unified data model. Once the data has been harmonized, it will be possible to publish it as a consistent dataset on the Semantic Web.

The critical component of the harmonization process is the deduplication and consolidation of actors, a challenging task due to the frequent occurrence of spelling variations, abbreviations, or errors in data. In historical material people can be referred to using multiple name variations: for example, Georg Zacharias Forsman was later known as Yrjö Sakari Yrjö-Koskinen. Furthermore, the existence of multiple individuals with identical names (e.g., cousins often had the same names) requires considering biographical information to disambiguate them. The completeness of such information varies between the data sources and actors, where we have plenty of information for prominent historical figures, but little for others.

In Europe, there exist numerous digital humanities initiatives that gather well-curated metadata (i.e., detailed information regarding senders, recipients, dates, and locations) from edited letter collections, including Europeana[1] (Doerr, 2010), ePistolarium[2] (Ravenek et al., 2017), SKILLNET[3], correspSearch (Dumont S. et al 2021), NorKorr (Rockenberger et al, 2019), and EMLO (Hotson et al., 2019). Although the mentioned projects provide valuable insights on actor deduplication, our task with unedited collections poses significant challenges as a substantial proportion of actors lack information beyond their correspondence. Thus, it is crucial for us to use available data on actors' social circle to develop accurate deduplication and reconciliation models. Our pioneering work presents unique challenges that have yet to be tackled in the field.

2. Data

The CoCo dataset primarily includes information about the sender, recipient, date of sending, and the number of letters exchanged. Supplementary details, such as the place of sending, language used, or people mentioned in the letter content, may also be provided. The dataset currently contains 359,246 letters and 39,884 actors, of which 6,376 (16%) have been linked to external sources.

To facilitate the information comparison process, we have two pools of actor information: the *CoCo Pool*, which comes from our source data, and the *CSV Pool*, which comes from external databases. These pools offer

[1]http://www.europeana.eu

[2]http://ckcc.huygens.knaw.nl/epistolarium/

[3]https://skillnet.nl

two main benefits: they save computing time by reducing the need for external database requests and allow us to take advantage of existing information comparison tools while optimizing our system development resources.

The CSV pool is generated to facilitate the disambiguation task, collecting external biographical information for historical actors from databases like Wikidata and Finnish actor ontologies like AcademySampo[4], BiographySampo[5], and Kanto[6]. This information includes the years of living, gender, place of birth, residence, death, variations of names, and IDs in external databases. The actors in the collection are limited to those living during the so-called *Long 19th century*, matching the main temporal period of correspondences in CoCo. Furthermore, when choosing a candidate entry from Wikidata, we prioritize actors with connections to Nordic countries, e.g., having entries in Nordic databases, or having lifetime activities or family relations related to Finland. The amount of data available for a single actor can vary between external sources, so the idea behind using an assembled database gathers and completes information from various available sources.

3. Deduplication Methods

The process of deduplication can be divided into a series of steps. First, we identify potential matching candidates. Next, we assign a confidence score to each potential match. Finally, we decide on whether to merge the records or not.

3.1 Candidate Matching

When it comes to candidate matching, we can compare candidates from the *CoCo Pool* and the *CSV Pool*. Since our data is limited in terms of actor information, matches with actors from the *CSV Pool* can help us identify duplicates. For example, if we have one person with two names in *CoCo Pool*, we can match each of the names with the same record in the *CSV Pool* to deduce that both names belong to the same person, as shown in Figure 1.

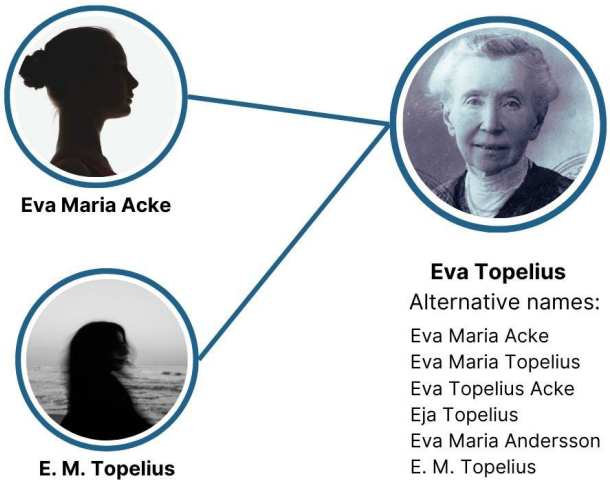

Figure 1: Eva Maria Acke and E. M. Topelius are the Same Person as They Both Correspond to an External Source Named Eva Topelius, Albeit Under Different Names. Eva's Picture Credit: Wikipedia

We can use various methods for candidate matching, with the most straightforward being name string matching and time matching.

String matching includes rules where we compare two strings and look for similarity. To accommodate expected irregularities, we incorporate a fuzzy string match by using the Jaro-Winkler score to determine the similarity of two strings. We define a threshold score for each rule to determine whether two names are the same. Another approach is described by Ulfves (2022), who has explored matching names by their phonetic

[4] https://akatemiasampo.fi/en/

[5] http://biografiasampo.fi/

[6] https://www.kiwi.fi/display/Toimijakuvailupalvelu/About+Kanto+in+English

similarity, which is particularly useful for historical records from a time when there was not standardized writing practice.

For time matching, we compare a time relative position to another time window. For example, we compare the start of active correspondence and the date of birth. In this rule, if the actor's active correspondence start time is earlier or too close to the actor's date of birth, then it is less likely that the match is accurate.

The final approach that we have explored to create matching candidates is a machine learning approach. We have trained word embeddings on the combined *CoCo Pool* and *CSV Pool* datasets. Then, we have computed cosine similarity of the name vectors, which gives us a distance score of different names in the vector space. This method works very well to identify potential matches; however, it gives wide results that need to be further filtered with other scoring methods.

3.2 Scoring

After identifying potential matches, it is crucial to score them to decide whether they are duplicate names of the same person or not. A higher matching score indicates greater accuracy in matching. We obtain confidence scores from the process of creating matching candidates using similarity measures, both for strings and vectors. Additionally, time matching can also provide a defined score based on overlap. However, to improve the system's accuracy, we've explored additional methods to score matches. These scores can be combined with the initial ones for a more comprehensive measure.

One of the additional scoring methods is the name frequency. Candidates with uncommon names are more likely to be the same person. Name frequency can be calculated from our datasets or historical person registers.

Another useful scoring method is derived from the correspondence network of first and second-degree connections. If candidates have not exchanged letters with each other or have overlapping social circles, there is a higher probability that they are the same person. For example, by examining the correspondence network in Figure 2, we can determine that Eljas Lönnrot (1822-1883) is not the same person as the famous Elias Lönnrot (1822-1884) since he sent a letter to him. On the other hand, Elias Lönnroth (bottom right) is likely the famous writer since he corresponded with members of the Fellman family, like the famous Elias. The fourth actor, Eljas Lönnrot (1895-1924), has a different correspondence circle and is a different person.

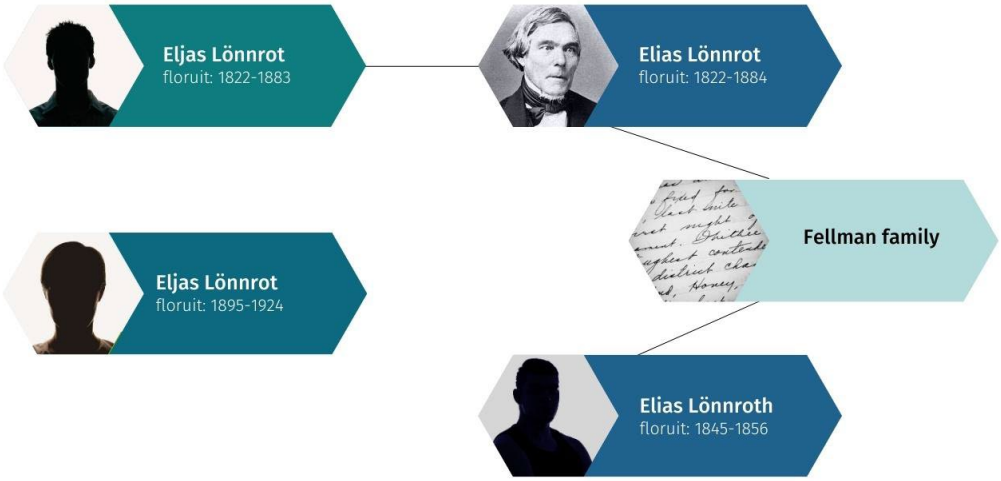

Figure 2: An Example of Four Actors who Share Similar or Identical Names and Have Matching Floruit Years. In This Scenario, the two Actors on the Right-Hand Side of the Image Were Identified as the Same Person. Elias' picture credit: Wikipedia

To get the social network scores, we have transformed correspondence metadata into vector space to measure the distance between persons who have conversed in similar social circles. The smaller the distance between matching candidates in vector space, the closer they were in the social circle.

Finally, by combining scores from different comparison methods, we obtain a scored list of deduplication candidates that can either be automatically accepted if the score is above a certain threshold or further manually checked by an expert.

3.3 Decision on Merging

To determine whether to combine two actors or not, we must rely on scores and a predefined threshold. If the level of confidence is remarkably high, the merging decision can be automated, or else a human expert can assess the matching candidates above a particular threshold and make a judgment. Alternatively, a neural classifier can be trained, which would utilize various scores and training data from prior manual decision-making to make informed predictions.

It is crucial to maintain a high level of accuracy while merging records, as well as to preserve provenance, which includes information about all the actions taken and the original data. Historians are the ultimate end-users, and it is vital to provide them with reliable information sources for their historical research. Consequently, the threshold for automated merging should be highly certain, and human decision-making must be cautious and precise, with no uncertain merging.

4. Discussion and Conclusions

This paper outlines the practices to create a strong and efficient process for removing duplicate person names. We employ a combination of string-matching algorithms, rule-based approaches, and machine learning techniques to generate a precise and scalable deduplication method.

Two critical aspects of our work are still under development. The first is how to utilize the scoring information to determine which match is accurate. While the intuitive approach is to sum all the scores and establish a threshold for the sum to qualify as a match, we need to explore alternative processing methods, such as a weighted scoring system, where certain scores are given greater importance.

The second issue is measuring the quality of the matching results. We must find an optimal balance between precision and completeness and evaluate the matching result accurately. At present, we rely on expert human evaluators to assess the results, which is time-consuming and limits the number of linkages evaluated. Additionally, we must re-evaluate results every time we modify the process, and this would be more convenient if we had a golden standard in place, which we are yet to develop.

Finally, to achieve maximum impact and facilitate the adoption of this deduplication method in future projects, it is crucial to develop a scalable and easily reusable method that can be easily applied to other datasets.

Acknowledgements

Our work was funded by the Academy of Finland as part of the project Constellations of Correspondence: Relational Study of Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland (CoCo) (decision numbers 339828, 340834, and 339918). CSC – IT Center for Science, Finland, provided computational resources for the work.

References

Doerr, Martin, et al. (2010) "The Europeana data model (EDM)." World Library and Information Congress: 76th IFLA general conference and assembly.

Drobac, Senka, et al. (2023) “The Laborious Cleaning: Acquiring and Transforming 19th-Century Epistolary Metadata”. *In review* (DHNB 2023).

Dumont S., et al. (2021) “correspsearch – connect scholarly editions of correspondence (2.0.0)”, Berlin–Brandenburg Academy of Sciences and Humanities, 2021. URL: https://correspSearch.net.

Hotson, H., Wallnig, T. (2019) “Reassembling the Republic of Letters in the Digital Age.” Göttingen University Press, https://doi.org/10.17875/gup2019-1146

Ravenek, Walter, Charles van den Heuvel, and Guido Gerritsen (2017) "The ePistolarium: origins and techniques." CLARIN in the Low Countries (2017): 317-323.

Rockenberger, Annika, et al. (2019) “Norwegian Correspondences and Linked Open Data.” Proceedings of the Digital Humanities in the Nordic Countries 4th Conference. CEUR Workshop Proceedings, 365-375.

Tuominen, Jouni, et al. (2022) “Constellations of Correspondence: A Linked Data Service and Portal for Studying Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland.” Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022).

Ulfves, Max. (2022) "Disambiguation of Named Entities in a Social Network."