Towards a Knowledge-Aware AI
A. Dimou et al. (Eds.)

doi:10.3233/SSW220012

Adding Domain Knowledge to Improve Entity Resolution in 17th and 18th Century Amsterdam Archival Records

J. Baas [a], L. van Wissen [b], J. Reinders [c], M. M. Dastani [a] and A. J. Feelders [a]
[a] *Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, Netherlands*
[b] *University of Amsterdam, Turfdraagsterpad 9, 1012 XT Amsterdam, Netherlands*
[c] *Huygens Institute for the History of the Netherlands, Oudezijds Achterburgwal 185, 1012 DK, Amsterdam, Netherlands*

Abstract. The problem of entity resolution is central in the field of Digital Humanities. It is also one of the major issues in the Golden Agents project, which aims at creating an infrastructure that enables researchers to search for patterns that span across decentralised knowledge graphs from cultural heritage institutes. To this end, we created a method to perform entity resolution on complex historical knowledge graphs. In previous work, we encoded and embedded the relevant (duplicate) entities in a vector space to derive similarities between them based on sharing a similar context in RDF graphs. In some cases, however, available domain knowledge or rational axioms can be applied to improve entity resolution performance. We show how domain knowledge and rational axioms relevant to the task at hand can be expressed as (probabilistic) rules, and how the information derived from rule application can be combined with quantitative information from the embedding. In this work, we perform our entity resolution method on two data sets. First, we apply it to a data set for which we have a detailed ground truth for validation. This experiment shows that the combination of embedding and the application of domain knowledge and rational axioms leads to improved resolution performance. Second, we perform a case study by applying our method to a larger data set for which there is no ground truth and where the outcome is subsequently validated by a domain expert. Results of this demonstrate that our method achieves a very high precision.

Keywords. Linked Open Data, Digital Humanities, Entity Resolution, Machine Learning, Embeddings

1. Introduction

The project *Golden Agents: Creative Industries and the Making of the Dutch Golden Age*[1] develops a research infrastructure to study relations and interactions between producers and consumers of creative goods during the 17th and 18th century in Amsterdam. It brings together heterogeneous data sets from several content providers as linked open data. However, the independent nature of the institutions that govern these data sets causes them to use different identifiers to refer to the same real-world object, if they pro-

[1] https://www.goldenagents.org

vide an identifier for these resources at all. Especially when dealing with archival data the situation is even more complicated, as it is rarely the case that entities are identified as more than a textual reference to e.g. a person or a location. Due to the size and type of this archival data, internal disambiguation or external reconciliation is often not available, which makes every reference an unresolved ambiguous one. In order to use these data sets for prosopographical (common characteristics of a group of people) and biographical research, and ask questions that shed more light on the production and consumption of cultural goods, we need a way to efficiently disambiguate these textual references to entities, thereby making it possible to understand the impact of the Dutch Golden Age on the creative industries and its actors.

The limited availability of ground truth in this type of data is a common problem in the field of the Digital Humanities, which makes it hard to apply supervised machine learning methods to the problem of entity resolution and entity linking. However, unsupervised machine learning methods such as embedding techniques can be applied to this type of data to create sub-symbolic models with which we can attempt to resolve these entity references. Nevertheless, the models may produce errors in the entity resolution outcome, where pairs of entities are identified as identical while in reality they are not. These errors may come about due to inherent weaknesses in the embedding technique, or incomplete and unreliable information in the original data. It is not viable to switch to a supervised method in an attempt to improve performance due to the lack of ground truth. Therefore, it is necessary to use other types of information, such as rational axioms and domain knowledge, to improve the entity resolution method.

We argue that domain-specific knowledge can be used to detect and correct errors and propose an approach to incorporate domain knowledge in an existing entity resolution algorithm. Examples of such domain-specific knowledge are (1) two entities cannot be identical if they occur in the same civil registration record, and (2) two entities, one with birth date x, the other with marriage date y cannot be identical if $x > y$. The purpose of these rules is to cast doubt on the conclusion of a sub-symbolic method that two entities are the same. Furthermore, the use of domain-specific knowledge can involve uncertainty due to the ambiguity in the data. For example, if we want to exploit the fact that a person cannot be born after they have died, one must be able to identify the born and dead persons unambiguously. The difficulty is that in this type of archival sources it is not a given that the person is actively involved in the registration event when they are mentioned. It could even be that the person is already deceased and is solely used as a disambiguating description for someone else (e.g. *Claartje Jans, widow of Cornelis Pieters*). If this happens in a burial registration or the registration of a testament, we need a rule that is aware of this knowledge and we do not treat the fact of a person's death as 100% certain, but see this as relative to the number of persons involved in the event. This uncertainty then propagates to the conclusion of the rule.

The novel contribution of this work is (1) an experiment with the incorporation of these rules with an existing embedding, and (2) a case study where we apply the method to more heterogeneous and voluminous data. In this paper we distinguish the terms 'experiment' and 'case study' to indicate that we have no sufficient ground truth available for the latter and that the main motivation for this case study was to apply the method to our data to make a linkset, which is then incorporated in the Golden Agents infrastructure. The purpose of the experiment is to test whether the inclusion of domain knowl-

edge improves entity resolution performance and to provide estimates for an optimal configuration of the method, which we use for the case study.

We show in our experiment that including the domain knowledge improves overall performance. We show with the case study, applied to a corpus of 84,268 resources, of which 22,073 (not fully disambiguated) persons participating in 7,339 different events, that our method proves to be a useful tool for entity resolution in heterogeneous archival data. We are able to disambiguate 9,151 entities with a precision of around 92%.

2. Related Work

Entity disambiguation on data from archives is a problem that has been around since the first digitisation initiatives of traditional (archival) indices and other entry points. These methods rely on either a strictly defined data model, or on the availability of sufficient data to disambiguate persons easily (e.g. a birth date). While the contribution of this work is on the integration of domain knowledge and rational axioms with embedded data, a selection of other methods for entity resolution on knowledge graphs are, Legato [1], LIMES [2], SILK [3], and Lenticular Lens [4]. These methods assume that there are only two data sets, often called source and target, and that these source and target data sets do not have internal duplicates. An exception is Lenticular Lens, which notes that it can work when the source and target are the same. Another difference is that in the first three methods only literals are used for disambiguation. Our method is more generic in the sense that all nodes related to an entity can act as context and it, in contrast to Lenticular Lens, does not solely rely on pre-defined rules to perform entity resolution. The more heterogeneous a data set is, the less feasible it is to work with a rule-based approach. These are major problems for applying other techniques to our data. Another difference is that our method takes into account the semantics of properties when comparing entities, something that is not done in, for instance, the bag-of-words model of Legato. This is important as we do not want to treat, for instance, death dates and birth dates equally in the context of an entity. That is, the fact that the birth date of an entity is the same as the death date of another entity does not mean they are similar.

In light of the use of integrating domain knowledge with embeddings, Guo et al. [5] have developed KALE, where an embedding is learned by jointly using facts from a knowledge graph and t-norm fuzzy logic. Similarly, Rocktäschel et al. [6] use first-order logic background knowledge to aid the matrix factorisation algorithm in learning dependencies between relations. Domain knowledge can be used with pre-trained embeddings as well, Wang et al. [7] predict new facts (also known as link prediction) by combining the output of an existing embedding with physical and logical rules into an integer linear programming problem. Others have worked on the very similar problem of author name disambiguation, where authors are linked in scholarly data sets such as dblp[2]. For instance, Cen et al.[8] learn stopping criteria to use with hierarchical agglomerative clustering. This method, however, requires a training set, which is often not or very limitedly available when working with cultural heritage data. Furthermore, hierarchical agglomerative clustering has the following drawbacks: its time complexity is high compared with some other clustering methods and it needs to take the number of clusters as input while

[2] https://dblp.org/

determining the number of clusters is usually an intractable problem or completely unknown in many digital humanities applications. Work on author name disambiguation seems to have died down, with some efforts ongoing in the medical domain, such as with Sanyal et al. [9].

Others have worked on similar cultural heritage data sets but without the use of embeddings. Raad et al. [10] create a certificate linking method for Dutch civil certificates from the Zeeland region, based on string similarity computations. Furthermore, they propose a contextual identity link [11], as they observe that the `owl:sameAs` link is often misused. Similarly, Idrissou et al. [12,13] have proposed a contextual identity link based on the use of related entities to construct evidence for likely duplicate pairs. An example of evidence is that two entities may co-occur in multiple records under similar names. Koho et al. [14] reconcile military historical persons in three registers. Hendriks et al. [15] use data from the Amsterdam Notarial Archives and Dutch East India Company (VOC) and perform both named entity recognition and record linkage. Finally, Efremova et al. [16,17] perform entity resolution on civil certificates by making use of name similarity (corrected for name popularity), proximity of locations, and limited co-occurrence information as features in regression tree and logistic regression models.

3. Data

The section below describes the data sets used in the experiment and the case study. Each of the data sets either already existed in RDF or was created in the Golden Agents project.

3.1. Amsterdam City Archives

For both the experiment and the case study, the collections of the City Archives of Amsterdam[3] play an important role. Traditionally, the Amsterdam City Archives served the interests of its users by unlocking its sources partially via paper indices that always contain a person's name and the date of the registration in the source. With the scanning of the original sources and the digitisation of the indices starting around 2000, it became easier to find the individuals mentioned in each index. The Golden Agents project has taken these digitised sources and converted and modelled them into RDF so that they are available to be connected to other data sets.

3.1.1. Notice of Marriage registrations

One of the most important indices for early modernists is the index on Notice of Marriage [=Ondertrouw] registrations. Whereas regular Marriage registrations do not contain that much information on the occupations, ages, residency, etc. of the groom and bride-to-be, the Notice of Marriage does contain this. This is crucial information for disambiguating persons. The index counts approximately 1 million person names and distinguishes roles (groom, bride) in the registration event. For 20% of this data, we also have information on among others the witnesses participating in the event. This data is coming from the crowdsourcing project *Ja, ik wil!* [=Yes I do!] [19]. Both the original Notice of Marriage index as well as this enrichment have been reconciled in the Golden Agents project.

[3] https://archief.amsterdam/indexen

```
@prefix pnv: <https://w3id.org/pnv#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix roar: <https://data.goldenagents.org/ontology/roar/> .
@prefix sem: <http://semanticweb.cs.vu.nl/2009/11/sem/> .
@prefix thes: <https://data.goldenagents.org/thesaurus/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix deed: <https://archief.amsterdam/indexen/deeds/a74a2bbf-20be-4c27-9d14-d477f5c09ea3?> .
@prefix personName: <https://data.goldenagents.org/datasets/personname/> .

deed:event=Event1 a thes:Begraven ;
    sem:hasTimeStamp "1789-10-24"^^xsd:date .

deed:person=99e87e23-d295-2bb2-e053-b784100a6a2e a roar:Person ;
    rdfs:label "Lucretia Wilhelmina van Merken" ;
    roar:participatesIn deed:event=Event1 ;
    pnv:hasName personName:5289fa40-1a30-5ac9-8583-6a4d01c6443a .

personName:5289fa40-1a30-5ac9-8583-6a4d01c6443a a pnv:PersonName ;
    pnv:baseSurname "Merken" ;
    pnv:givenName "Lucretia Wilhelmina" ;
    pnv:literalName "Lucretia Wilhelmina van Merken" ;
    pnv:surnamePrefix "van" .

deed:person=99e87e23-d294-2bb2-e053-b784100a6a2e a roar:Person ;
    rdfs:label "Nicolaas Simon van Winter" ;
    roar:participatesIn deed:event=Event1 ;
    pnv:hasName personName:b3f66e9f-9f64-5d45-bbe0-24343cd3a90f .

personName:b3f66e9f-9f64-5d45-bbe0-24343cd3a90f a pnv:PersonName ;
    pnv:baseSurname "Winter" ;
    pnv:givenName "Nicolaas Simon" ;
    roar:carriedBy deed:person=99e87e23-d295-2bb2-e053-b784100a6a2e ;
    pnv:literalName "Nicolaas Simon van Winter" ;
    pnv:surnamePrefix "van" .

_:role1 a thes:Geregistreerde ;
    roar:carriedIn deed:event=Event1 .

_:role2 a thes:Geregistreerde ;
    roar:carriedBy deed:person=99e87e23-d294-2bb2-e053-b784100a6a2e ;
    roar:carriedIn deed:event=Event1 .
```

Listing 1: Example RDF Turtle syntax for a single event (Burial registration) in the subset in which two persons participate in the role of 'being registered'. Their name is described as a separate resource using the Person Name Vocabulary (PNV) [18].

3.1.2. Baptism registrations

Containing almost 5 million person names, the index on the Baptism registrations is the largest early modern one the City Archives of Amsterdam has. As does the Notice of Marriage index, the index on the Baptisms includes the roles in which persons are mentioned: child, father, mother, and sometimes witness, if available in the source. Also, the churches where the event took place, are mentioned. This index is highly interesting for those looking for networks of family, friends and of religious congregations.

3.1.3. Burial registrations

One of the oldest indices of the Amsterdam City Archives is the one made on the burial registrations. Although this index is quite a large one, as it contains approximately 1.5 million person names, it cannot be considered trustworthy due to its provenance. For instance, it is unclear which of the mentioned persons is buried, and which person is mentioned as partner, serving as disambiguating description or being the one that declared the burial event. For this reason, the only role persons have in the burial event is the one of 'being registered', which is problematic for the application of rules. Next to this, not all the registration books of burials in churches and on graveyards in Amsterdam survived

over time. Although useful as an entrance to the source, one has to be careful in drawing final conclusions in (not) matching records from the burial records with other sources.

3.1.4. Notarial archives

Contrary to the former three mentioned indices, no digitised index was available of the Amsterdam notarial archive up to 2016. Together with the start of the Golden Agents project and for a large part co-financed by it, the City Archives started in 2016 with the indexing of its largest early modern archival collection. From this 3,5 kilometres plank length counting archive almost all early modern deeds have been scanned, delivering nearly 10 million scans. Approximately two million scans of them have been indexed on deed type, person names, date, and locations in the crowdsourcing project 'Alle Amsterdamse Akten' [=All Amsterdam Deeds],[4] resulting in almost 600k unique deeds with over 3 million person names. In this study, three deed types have our special attention: Last Wills [=Testamenten], Prenuptial Agreements [=Huwelijkse Voorwaarden] and Probate Inventories [=Boedelinventarissen]. These deeds are highly interesting because, contrary to the other mentioned indices, each deed gives insight into a complete network of family and friends. Unfortunately, the only role of people in the registration event that is included in the index is the one of 'being registered'. The records in this index resemble the burial registration records in this respect.

3.2. Golden Agents / University of Amsterdam

ECARTICO[5] is a comprehensive collection of biographical data from cultural entrepreneurs, such as painters, writers, book printers, illustrators, goldsmiths and related figures from ca. 1475-1725 in The Netherlands. It aggregates information on persons from this time period and includes references to both primary and secondary sources, among others the sources of the Amsterdam City Archives. The data set thereby provides a basis for the creation of a ground truth for person disambiguation. ECARTICO is hosted by the University of Amsterdam and published as Linked Open Data in the SCHEMA.ORG vocabulary. It is included in the Golden Agents project as one of the main data sets.

3.3. Golden Agents / KB, the National Library of The Netherlands

This data set with Occasional Poetry published in the Dutch Republic between ca. 1600-1800 built by the Royal Library of the Netherlands (KB) contains 6,906 printed poems or collections of poems on a particular type of event, such as marriage (2,433), death (1,049), or various types of anniversaries (474). Work on the data set has been concluded by the KB,[6] and the data set was converted to linked data in the Golden Agents project. It holds bibliographical information on a poem written for an event, together with information on the event's date and participants.

We disambiguated textual references to authors and printers/publishers and connected them to existing resources for metadata: the Dutch Thesaurus of Author names

[4] https://alleamsterdamseakten.nl/

[5] https://www.vondel.humanities.uva.nl/ecartico/

[6] https://www.kb.nl/bronnen-zoekwijzers/databanken-mede-gemaakt-door-de-kb/gelegenheidsgedichten-tot-1800-in-nederland

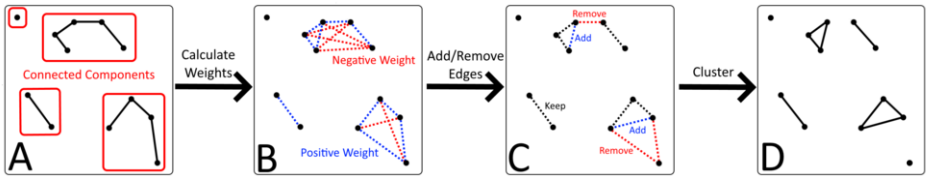

Figure 1. An overview of the entity resolution process.

(NTA)[7] and the STCN printer thesaurus[8] respectively. The same was done, if possible, for persons that are mentioned in relation to the event that the poem is written on, such as the bride and groom in a marriage. In the case that these author and person references could not be found in an existing thesaurus (e.g. the NTA or Wikidata[9]), we tried to find mentions of these actors in archival sources in the above-described data sets of the Amsterdam City Archives and connected them accordingly by making use of a rule-based linking approach using the Lenticular Lens tool [4].

4. Method

We extend the method for identifying duplicate entities previously presented in [20] of which figure 1 gives an overview. This method makes use of an embedding, that is, in our case, an n-dimensional Euclidean space where each (non-unique) person resource is assigned a coordinate based on its neighbouring nodes in the RDF graph. In this research, these neighbouring nodes can be other person entities, events, or names. We call these neighbouring nodes the *context* of a *focus* node. Entities in a context are weighted according to their proximity to the focus node and the number of possible routes from the focus node to a context node in the RDF graph. More detailed information on how this embedding is created can be found in our previous work [21,20].

Following this method, we start with a set of entities that are represented as a (sub-symbolic) embedding $\mathscr{E} = \{v_1, \ldots, v_m\}$ where each vector $v \in \mathscr{E}$ corresponds to a (non-unique) entity (e.g. a person resource in our data set) which may have duplicates (i.e. other resources that refer to the same real-life object). We construct a graph $G = (V, E)$ where each vertex in V corresponds to a vector in $\mathscr{E}$. We then take the k (approximate) nearest neighbours based on euclidean distance between embedding vector of each entity $i \in V$, denoted by the set N_i^k, and create an edge between that entity and its neighbour $j \in N_i^k$ if their cosine similarity $u_{ij} = cosim(v_i, v_j)$ exceeds some threshold θ. Such constructed graphs, which are illustrated in panel A on figure 1, consist of a number of connected components. The number and size of these components depend on the choice of θ: high values of $\theta \approx 1$ result in many small components and lower values result in fewer but larger components. With each possible pair of entities i and j within each component we associate a weight $w_{ij} = u_{ij} - \theta$. Panel B illustrates how these weights can be both positive (blue, similar) and negative (red, dissimilar). Components are subsequently subdivided into cliques using integer linear programming (ILP) and an alternative heuris-

[7] http://data.bibliotheken.nl/id/dataset/persons
[8] http://data.bibliotheken.nl/id/dataset/stcn/printers
[9] https://www.wikidata.org/

tic algorithm (panel C). Both algorithms work by associating a cost for omitting a pair with a positive weight and for retaining a pair with a negative weight in a solution. The partitioning with the (approximate) lowest cost is then selected, illustrated in panel D.

In the remainder of this section, we explain our extension to the work described above. That is, how domain knowledge is encoded and integrated into this process at two points. This domain knowledge is encoded as a set of rules provided by domain experts. These rules, which aim at identifying non-duplicate entities, have the general form: *if a certain condition holds for a pair of entities i and j, then it is ruled out that i and j are duplicates*. We assume that we have a data set that contains additional knowledge about the entities on which the condition of these rules can be checked. This data can, for instance, come in the form of an RDF graph, where rules are encoded as filter clauses in SPARQL queries. We have provided an example SPARQL query in our repository[10]. Such rules can be divided into two categories:

1. **Definite Rules:** There are rules for which we know in advance that applying them results in conclusions with high certainty, that is, the two entities on which the definite rule is applied are not duplicates. For example, *if entity i and j both occur in the same marriage event, then it is ruled out that entities i and j are duplicates* as one can not be a bride/groom and witness at the same time.
2. **Probabilistic Rules:** For other rules, we may not be as confident in the evaluation of the premise. This can, for instance, be due to uncertainty or ambiguity in the data. For example, consider the rule *if the burial date of entity i is before the marriage date of entity j, then it is ruled out that entities i and j are duplicates*. The burial record is ambiguous as it contains in addition to entity i also an entity k, without mentioning who was buried. This means that the probability that entity i was the one who died is 50%, and the probability of the conclusion of the rule that entities i and j are ruled out as duplicates also becomes 50%.

We use $\mathscr{S}_{ij}$ to denote the set of all definite rules applied to i and j, and $\mathscr{R}_{ij} = \{r_1, \ldots, r_n\}$ to denote the set of all probabilistic rules that are applied to entities i and j. First, the definite rules are used to further cull the approximate nearest neighbours found in the embedding, next to already removing neighbours that have a similarity which falls below the threshold θ. That is, we only create an edge between entities i and j in the graph G if $u_{ij} \geq \theta$ and no definite rule is satisfied for that pair. Since multiple probabilistic rules could be satisfied on an entity pair, the rules and their probabilities need to be aggregated first. To this end, for $r \in \mathscr{R}_{ij}$, we use $p(r)$ to denote the probability that i and j are ruled out as duplicates, e.g. $p(r) = 0.5$ is read as in 50% of the cases it is ruled out that i and j are duplicates. When evaluating a set of rules on a pair of entities i and j, the outcomes are aggregated as follows:

$$p_{ij} = \prod_{r \in \mathscr{R}_{ij}} 1 - p(r), \tag{1}$$

where p_{ij} is the total probability that it is *not* ruled out that i and j are duplicates.

[10] https://github.com/knaw-huc/golden-agents-occasional-poetry

Here we assume for convenience that rules are independent, as it would be very difficult to quantify the dependencies between all of them. As is, for example, the case with the naive Bayes classifier, we expect that the independence assumption still gives a reasonable approximation.

Using equation (1), we calculate a penalty s_{ij} for considering a given pair of entities i and j as duplicates:

$$s_{ij} = \begin{cases} 10^6 & \text{if } \mathscr{S}_{ij} \neq \emptyset \\ 0 & \text{else if } \mathscr{R}_{ij} = \emptyset \\ 1 - \sqrt{p_{ij}} & \text{otherwise} \end{cases} \tag{2}$$

This yields a very large penalty if any definite rule was applied for the pair of entities i and j and no penalty if no rule was applied. This large penalty guarantees that pairs of entities satisfy a definite rule in the final results. In all other cases, it transitions smoothly between $s_{ij} = 1$ for $p_{ij} = 0$, and $s_{ij} = 0$ for $p_{ij} = 1$. The square root is taken to reduce the penalty to prevent the probabilistic rules from acting as definite rules in some edge cases.

Finally, we calculate the updated final weight w_{ij}, used for the partitioning of connected components, with the following equation:

$$w_{ij} = u_{ij} - s_{ij} - \theta \tag{3}$$

Note that s_{ij} represents the amount of evidence against the conclusion that i and j are duplicates. Also, note that rule application can never lead to an increase in the weight since a lack of evidence to rule out a pair of entities as duplicates do not constitute evidence that they are the same. At this point, the connected components are then partitioned into cliques as in our previous work. Each clique then corresponds to a unique real-life object (panel D). For example, in our experiment, each vertex (entity) denotes a *reference* to a person, and each clique corresponds to a single real-life person (object).

5. Experimental Setup

For both the experiment and the case study, we use domain knowledge to improve the results. The experiment shows by how much the rules improve performance in both precision and recall, while the case study only gives us precision. The experiment makes it possible to determine reasonable values of θ for the case study when both precision and recall are considered.

5.1. Domain Knowledge Experiment

For our experiment to determine the effectiveness of applying rules we have used four data sets containing real historical data from the cultural heritage domain. We made use of a subset of the above-described registers from the Amsterdam City Archives and a subset of the ECARTICO data set containing information on marriage registrations. Combined, the above-described subsets together form a data set that contains 12,517

entities referring to (non-unique) persons, and a partial ground truth of 1073 clusters, obtained with manual validation by domain experts [22].

We have designed two definite rules based on expert domain knowledge, namely that two entities co-occurring in the same record are not the same, and that two entities originating from ECARTICO (section 3.2) are not the same. The latter rule is based on the knowledge that ECARTICO is an expert-curated biographical data set that contains a single unique entry for a single person. Furthermore, we have constructed 8 probabilistic rules that consider the dates at which the events took place, and how many years occur between them. Below we list three example rules for entities i and j:

1. Entity i is mentioned in the role of groom or bride in a notice of marriage registration at date d_1, and j is mentioned in the role of husband, wife, father, or mother in a marriage or baptism registration respectively at date d_2. If more than 30 years occur between d_1 and d_2, then it is ruled out that i and j are duplicates.
2. i is mentioned in the role of groom or bride at date d_1. If there is less than 17 years between the birth date of j and d_1, then it is ruled out that i and j are duplicates.
3. i is mentioned in the role of groom or bride at date d_1. If there is more than 60 years between the birth date of j and d_1, then it is ruled out that i and j are duplicates.

Note that in many cases, a rule does not apply because one or both attributes are missing for a given pair of persons. All rules are combinations of what the dates represent and how they relate to each other. The rules that contain a burial date in the premise have been given a probability of 0.5 for reasons explained in section 3.1.3. All other rules have been assigned a range of probabilities ranging from 0.25 to 0.95, with the justification that, for cultural reasons, these kinds of rules can be trusted with different levels to be correct in their assessment. We test our method four times, each time with a different set of rules: (a) no rules, (b) only definite rules, (c) only probabilistic rules, and (d) all rules.

We apply a combination of the Correlation Clustering [23] and Vote [24] algorithms, as these algorithms performed best in our previous work [20]. Due to the NP-hardness of the Correlation Clustering (ILP) algorithm, we apply it only to connected components smaller than 100 vertices. Larger components are handled with the Vote algorithm. This is mostly an issue when using low values of θ, where large components can be generated. As θ increases, Correlation Clustering is used more, until it is solely used when $\theta \geq 0.70$. We have shown in previous work [25] that in our setting the Vote algorithm is very competitive with Correlation Clustering. These clusters are then compared to a partial ground truth which has been validated by domain experts.

5.2. Occasional Poetry Case Study

In the case study we focus on combining two data sets that come from two content providers: (1) the indices from the Amsterdam City Archives, and (2) the Occasional Poetry data. The latter is used to make a selection of these data sets by which we create a subset that is both relevant for our case study and is workable in terms of size.

We limit the data to resources related to the actors in the Occasional Poetry data set, in particular only the persons that we reconciled with at least one mention in the Amsterdam City Archives data. We create this subset using a SPARQL query that constructs a copy of the data that only includes events in which at least one disambiguated person was mentioned. For each of these events, it gives the date of the event, the type of event,

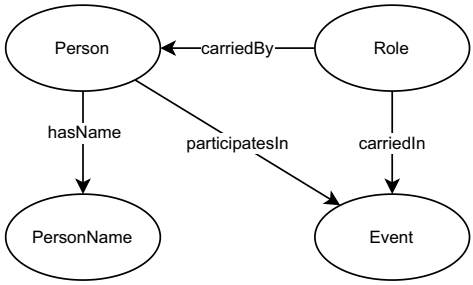

Figure 2. Basic structure of our RDF-model. Persons participate (either actively or passively) in an event. In the event, they carry a specific role: the role in which they are mentioned. Every person has one or more names.

the persons participating in the event, and the role in which they participate (e.g. bride, or witness). In total, this produces a data set that contains 7,339 events and 22,073 persons, of whom 3,839 have been disambiguated (i.e. they participate in at least two events: one from the Occasional Poetry data set, and one from the City Archives' data sets). An example of resources involved in a single event can be seen in Figure 1. The resources are following a basic format that models resources as part of an event in a particular role. The THES: prefixed classes refer to a thesaurus of event and role types. The SPARQL query and the subset itself can be seen in our documentation on GitHub.[11]

In this case study, our goal is to maximise the disambiguation of entities in our subset, so that, ideally, a single (disambiguated) person participates in more than one event. With this, we will be able to gain insight into the lives of persons in this data set, such as their lifespan and their social or professional network, and the motivation behind creative production in the 17th and 18th century in Amsterdam. As mentioned under section 3.3, we already made connections from the Occasional Poetry data set to the data sets of the Amsterdam City Archives using the Lenticular Lens tool [4]. These links were all manually validated and allow us to create the subset described under section 5.2.

Then we extend these links by making use of the graph embeddings method to in particular disambiguate peripheral entities. These are most often entities that do not occur in the role of bride, groom, father, or mother, and that only are mentioned in the Amsterdam City Archives data set. We use three definite rules and no probabilistic rules in the case study. Entities i and j are considered not to be duplicates if any of the following conditions hold:

1. Both i and j participate in the same event.
2. Both i and j are mentioned in the role of child in a baptism record, assuming there are no duplicate baptism records.
3. If i is mentioned in the role of a child in a baptism record and j is mentioned anywhere else in any role at an earlier date.

Finally, a domain expert validates the results of the embedding to assess the quality of the results as well as their usability for future research. This is done by indicating whether or not a person resource (represented by a contextualised URI) refers to the same entity as other URIs in the same cluster. Extra metadata is given to ease the validation process,

[11] https://github.com/knaw-huc/golden-agents-occasional-poetry

	Precision						Recall					
θ	0.60	0.65	0.70	0.75	0.80	0.85	0.60	0.65	0.70	0.75	0.80	0.85
previous work	0.65	0.74	0.82	0.86	0.91	0.93	0.69	0.66	0.62	0.58	0.52	0.42
definite	0.87	0.90	0.93	0.93	0.94	0.95	0.68	0.66	0.61	0.57	0.51	0.42
probabilistic	0.69	0.77	0.85	0.89	0.92	0.94	0.68	0.65	0.61	0.57	0.51	0.42
all rules	0.87	0.91	0.93	0.93	0.94	0.95	0.68	0.65	0.61	0.57	0.51	0.42
	F1-Score						$F\frac{1}{2}$-Score					
θ	0.60	0.65	0.70	0.75	0.80	0.85	0.60	0.65	0.70	0.75	0.80	0.85
previous work	0.67	0.70	0.71	0.69	0.66	0.58	0.66	0.72	0.77	0.79	0.79	0.75
definite	0.76	0.76	0.74	0.71	0.66	0.58	0.82	0.84	0.84	0.83	0.81	0.76
probabilistic	0.68	0.70	0.71	0.69	0.65	0.58	0.69	0.74	0.79	0.80	0.79	0.75
all rules	0.76	0.76	0.74	0.71	0.66	0.58	0.83	0.84	0.84	0.83	0.81	0.76

Table 1. Precision, recall, F1 and $F\frac{1}{2}$-Scores for a range of different θ values and rule sets.

such as a person's name, their role, and the type of event they participate in. If necessary, the expert can inspect a scan of the original handwritten document.

Though not needed for the scoring of the method, we do generate an RDF linkset from the result for usage in the Golden Agents project which can be found, together with the data, in the repository.[12]

6. Results

Since we have a ground truth for the experiment on domain knowledge, we report both precision and recall. Then we use the F-Score to determine optimal values for θ. This optimal value is used again in the case study.

6.1. Results of the Domain Knowledge Experiment

Table 1 shows the results of our experiment. High θ values will produce very high precision, as many small (or even singleton) connected components are created. Each component will likely be composed of pairs with similarly high weights, suggesting a high likelihood of them referring to the same real-life object. On the other hand, high θ values will yield a low recall, as we exclude many pairs of entities with lower (but still reasonable) cosine similarities. We discuss each performance metric in more detail below.

6.1.1. Precision

From table 1 we can see that precision increases for all three combinations of rules. For very high values of θ, precision does not increase as the rules are unlikely to exclude pairs with very high cosine similarities, showing that the embedding correctly captured (part of) the likeness between entities. When we lower θ to more reasonable values, the rules start to exclude some of the pairs of entities which the embedding encoded as likely but not certainly referring to the same entity.

[12] https://github.com/knaw-huc/golden-agents-occasional-poetry

6.1.2. Recall

Table 1 shows that there is a very slight decrease in recall when rules are applied. We surmise that the decreases in recall are caused by a conflict between the data and the judgement of the domain experts who created the ground truth. For instance, some rules can cause false-negative pairs to be created when a domain expert previously judged a pair to be correct, even though it conflicts with a rule, thereby decreasing recall. This is usually caused by errors or uncertainty in the data, which is not uncommon in these kinds of corpora.

6.1.3. F-Scores

The F-Score can be used to pick an appropriate value for θ based on both precision and recall. It is possible to adjust the F-Score to give greater weight to either precision or recall, depending on the situation. In our case, we report both F1 and $F\frac{1}{2}$-Score, which weights precision twice as high as recall, as in our particular case precision is more important than recall. This is due to the fact that errors in the entity resolution process can complicate the further analysis of the data by historians. A threshold of $\theta = 0.70$ (i.e. the score with the highest $F\frac{1}{2}$-Score) gives the optimum result when precision is valued twice as high as recall, otherwise the threshold $\theta = 0.65$ is best.

6.2. Results of the Occasional Poetry Case Study

We based the choice of $\theta = 0.70$ (as explained in section 5) on the outcome of the experiment, also taking into consideration that this rendered a result size that can be validated in a reasonable time (within a week) by a domain expert. This resulted in $9,151$ entities (references to persons) being clustered into $3,400$ distinct clusters, where each cluster represents the occurrences of a single real-life person. Of these $9,151$ entities, $8,326$ were correctly clustered according to the domain expert. Furthermore, 45 entities could not be confidently attributed as either correct or incorrect due to the lack of available information needed for the validation. This means that we achieve a precision between 0.91, if all 45 entities are incorrect, and 0.92 if they are all correct. We are unable to compute the recall, as it is not feasible due to the size, scope, and selection process of the data to determine, for each cluster, which entities are missing. However, we are able to present a list of common errors and points of improvement that were found during the validation process:

- In some cases, the logical ordering of baptism in the role of child, then notice of marriage as husband or wife, then baptism again in the role of father or mother could have been taken into account. Nonetheless, it should be said that people could and did remarry, so strict enforcement of a rule that states that baptism always takes place after a marriage, could introduce additional errors while removing others.
- Some entities were wrongfully clustered. Adding additional relations to other nodes (e.g. religion types, church information) in combination with formulating extra rules could have been used to refine the disambiguation.
- The common occurrence of some patrician family names causes them to be put together into a single cluster, e.g. in the case of the name ‘Jan Six’ that appears in every generation of this family.

- In some cases, two entities with very different names were put into the same cluster. This type of error can occur when two entities are clustered together with a similarity higher than the threshold θ. It is relatively simple to remove these errors with an additional string similarity check. This is something we plan for future work.

7. Conclusions

We have shown that it is possible to combine symbolic and sub-symbolic knowledge in such a way that it improves the performance of an existing entity resolution method. However, the interaction between the two is not always obvious. Introducing rules which, at first glance, should always improve performance may, in fact, worsen performance if there are errors in the data. In future work we plan on including positive evidence as well, that is, the inclusion of rules which, if applicable, indicate that entities are more likely to be duplicates. This can be done with logical rules, as well as by including information from symbolic methods such as [12].

The case study shows that the method works on a larger data set as well and that it gives good results. In future work, we plan to include more contextual information and distinctive attributes to the resources in the graph, such as religions and locations, which potentially improves the entity resolution outcome. Additionally, having a data set with thousands of disambiguated interconnected entities makes it possible to perform community detection. This could yield previously hidden patterns such as larger networks of people who were somehow connected to each other, either by social or professional relation. In particular, our method of entity resolution shows promise to work well in other deed types of the Amsterdam Notarial Archives, as only contextual information such as co-occurrence of people is available. In fact, this method can be applied to data from other archives in The Netherlands, Flanders, South-Africa, and possibly Suriname and Indonesia as well, as they share a similar structure and carry the same characteristics. Our generic method that can be applied to any (RDF) graph database, combined with tailored domain-rules, makes this tool highly applicable to the domain of cultural heritage.

Finally, We give extra attention to the issue of reproducibility, therefore we have published all the files necessary to run the experiment and case study in a git repository.[13]

References

[1] Achichi M, Bellahsene Z, Ellefi MB, Todorov K. Linking and disambiguating entities across heterogeneous RDF graphs. Journal of Web Semantics. 2019;55:108-21.

[2] Ngomo ACN, Auer S. LIMES—a time-efficient approach for large-scale link discovery on the web of data. In: Twenty-Second International Joint Conference on Artificial Intelligence; 2011. .

[3] Jentzsch A, Isele R, Bizer C. Silk-generating rdf links while publishing or consuming linked data. In: 9Th international semantic web conference (ISWC'10). Citeseer; 2010. .

[4] Idrissou A, Van Wissen L, Zamborlini V. The Lenticular Lens: Addressing Various Aspects of Entity Disambiguation in the Semantic Web; 2022. Graphs and Networks in the Humanities 2022, 3-4 February. Amsterdam, The Netherlands.

[13] https://github.com/knaw-huc/golden-agents-occasional-poetry

[5] Guo S, Wang Q, Wang L, Wang B, Guo L. Jointly embedding knowledge graphs and logical rules. In: Proceedings of the 2016 conference on empirical methods in natural language processing; 2016. p. 192-202.
[6] Rocktäschel T, Singh S, Riedel S. Injecting logical background knowledge into embeddings for relation extraction. In: Proceedings of the 2015 conference of the north American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2015. p. 1119-29.
[7] Wang Q, Wang B, Guo L. Knowledge base completion using embeddings and rules. In: Twenty-fourth international joint conference on artificial intelligence; 2015. .
[8] Cen L, Dragut EC, Si L, Ouzzani M. Author disambiguation by hierarchical agglomerative clustering with adaptive stopping criterion. In: Proceedings of the 36th International ACM SIGIR conference on Research and development in information retrieval; 2013. p. 741-4.
[9] Sanyal DK, Bhowmick PK, Das PP. A review of author name disambiguation techniques for the PubMed bibliographic database. Journal of Information Science. 2021;47(2):227-54.
[10] Raad J, Mourits R, Rijpma A, Schalk R, Zijdeman R, Mandemakers K, et al. Linking Dutch civil certificates. In: Adamou A, Daga E, Meroño-Peñuela A, editors. WHiSe 2020 Workshop on Humanities in the Semantic Web 2020. CEUR Workshop Proceedings. CEUR-WS; 2020. p. 47-58. 3rd Workshop on Humanities in the Semantic Web, WHiSe 2020 ; Conference date: 02-06-2020.
[11] Raad J, Pernelle N, Saïs F. Detection of contextual identity links in a knowledge base. In: Proceedings of the knowledge capture conference; 2017. p. 1-8.
[12] Idrissou AK, Hoekstra R, Van Harmelen F, Khalili A, Van Den Besselaar P. Is my:sameAs the same as your:sameAs? Lenticular Lenses for context-specific identity. In: Proceedings of the Knowledge Capture Conference; 2017. p. 1-8.
[13] Idrissou A, Zamborlini V, Van Harmelen F, Latronico C. Contextual entity disambiguation in domains with weak identity criteria: Disambiguating golden age amsterdamers. In: Proceedings of the 10th International Conference on Knowledge Capture; 2019. p. 259-62.
[14] Koho M, Leskinen P, Hyvönen E. Integrating historical person registers as linked open data in the warsampo knowledge graph. Semantic Systems In the Era of Knowledge Graphs SEMANTiCS. 2020:118-26.
[15] Hendriks B, Groth P, van Erp M. Recognising and Linking Entities in Old Dutch Text: A Case Study on VOC Notary Records. In: Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age; 2021. p. 25-36.
[16] Efremova I. Mining social structures from genealogical data [PhD thesis]. School of Mathematics and Computer Science Technische Universiteit Eindhoven; 2016.
[17] Efremova J, Ranjbar-Sahraei B, Rahmani H, Oliehoek FA, Calders T, Tuyls K, et al. Multi-source entity resolution for genealogical data. In: Population reconstruction. Springer; 2015. p. 129-54.
[18] Petram L, Dechesne E, Kruithof G. Person Name Vocabulary; 2019. Version 1.1. Available from: https://w3id.org/pnv.
[19] Van Weeren R, De Moor T. Counting Couples: The Marriage Banns Registers of the City of Amsterdam, 1580–1810: Social and Economic History. Research Data Journal for the Humanities and Social Sciences. 2021;6(1):1 45.
[20] Baas J, Dastani MM, Feelders AJ. Entity Matching in Digital Humanities Knowledge Graphs. In: Ehrmann M, Karsdorp F, Wevers M, , Andrews TL, Burghardt M, et al., editors. Proceedings of the Conference on Computational Humanities Research 2021. No. 2989 in CEUR Workshop Proceedings. Amsterdam, the Netherlands; 2021. p. 1-15.
[21] Baas J, Dastani M, Feelders A. Tailored graph embeddings for entity alignment on historical data. In: Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services; 2020. p. 125-33.
[22] Idrissou A, Zamborlini V, Latronico C, van Harmelen F, van den Heuvel C. Amsterdamers from the Golden Age to the Information Age via Lenticular Lenses; 2018. Presented at DHBenelux 2018, 6-8 June. Amsterdam.
[23] Bansal N, Blum A, Chawla S. Correlation clustering. Machine learning. 2004;56(1):89-113.
[24] Elsner M, Schudy W. Bounding and comparing methods for correlation clustering beyond ILP. In: Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing; 2009. p. 19-27.
[25] Baas J, Dastani MM, Feelders AJ. Exploiting Transitivity for Entity Matching. In: European Semantic Web Conference. Springer; 2021. p. 109-14.