Contextual entity disambiguation in domains with weak identity criteria

Idrissou, Al; Zamborlini, Veruska; Van Harmelen, Frank; Latronico, Chiara

published in
K-CAP 2019
2019

DOI (link to publisher)
10.1145/3360901.3364440

document version
Publisher's PDF, also known as Version of record

Contextual Entity Disambiguation in Domains with Weak Identity Criteria: Disambiguating Golden Age Amsterdamers

Al Idrissou
Frank van Harmelen
{o.a.k.idrissou,frank.van.harmelen}@vu.nl
Department of Computer Science,
Vrije Universiteit Amsterdam
Amsterdam, Netherlands

Veruska Zamborlini
Chiara Latronico
{v.zamborlini,c.latronico}@uva.nl
Institute for Logic, Language and Computation,
University of Amsterdam
Amsterdam, Netherlands

ABSTRACT

Entity disambiguation is a widely investigated topic, and many matching algorithms have been proposed. However, this task has not yet been satisfactorily addressed when the domain of interest provides poor or incomplete data with little discriminating power. In these cases, the use of content fields such as name and date is not enough and the simple use of relations with other entities is not of much help when these related entities also need disambiguation before they can be used. Therefore, we propose an approach for the disambiguation of clustered resources using *context* (related entities that are also clustered) as evidence for reconciling matched entities. We test the proposed method on datasets of historical records from Amsterdam in the 17th century for which context is available, and we compare the results of the proposed approach to a gold standard generated by three experts, which we make available online. The results show that the proposed approach manages to meaningfully use context for isolating identity sub-clusters with higher quality by eliminating potentially false positive matches.

CCS CONCEPTS

• **Information systems** → **Decision support systems**; • **Applied computing** → Arts and humanities.

KEYWORDS

entity disambiguation, entity reconciliation, entity resolution, data integration, linked data

ACM Reference Format:
Al Idrissou, Frank van Harmelen, Veruska Zamborlini, and Chiara Latronico. 2019. Contextual Entity Disambiguation in Domains with Weak Identity Criteria: Disambiguating Golden Age Amsterdamers. In *Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19), November 19–21, 2019, Marina Del Rey, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3360901.3364440

K-CAP '19, November 19–21, 2019, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-7008-0/19/11…$15.00
https://doi.org/10.1145/3360901.3364440

1 INTRODUCTION

Disambiguation is an entity matching process to determine if a pair of entities is identical or not. This process typically relies on specific properties that ascertain when two entities are the same or different. Irrespective of the algorithm used (e.g. AML [3], I-MATCH [7], LEGATO [1], LOGMAP [6], NJULINK [8] or any other), and irrespective of whether these algorithms are rule-based ([1, 3, 6–8]) or based on machine learning [11], they all implicitly or explicitly try to detect such distinctive properties and use them as the basis for the matching process. Since such discriminating information is essential for data integration, data owners do their utmost to include it in their data. Successful examples include: the citizen service number, the postal code system, the digital object identifier, the digital author identifier, the open researcher and contributor identifier and many more. However, such identifiers or identity criteria are not always explicitly available in real-life datasets. In such cases, entity matching algorithms end up using low power identity criteria for link discovery, such as names of persons, or their date of birth, leading these algorithms to include significant noise (false positives) in their resulting Identity Link Network (ILN) candidates space. Clearly, the inclusion of just one false positive node in a perfect ILN certainly results in a wrong ILN. Approaches such as [2, 10] compensate for the lack of sufficient identity criteria in the data at hand by combining a number of potentially weak atomic attribute values in the quest of matching with the ultimate goal to disambiguate matched entities. However, even by doing so, many ILNs are still corrupted with false positives. This motivates the problem tackled in this paper: *how to remove noise from ILNs that have been constructed on the basis of weak identity criteria*?

One way to remove false positive caused by properties with low discriminating power is suggested in Section 3, namely to rely on the context of *associated networks*, where *an associated network is a network that contains entities that participate in shared events*. Such associations can then be used to dynamically adjust our trust in a particular identity link: the more such associated networks are interlinked, the higher the trust in a particular identity network.

The quality of such graphs is of crucial importance as it is essential for data integration. For that, we contribute with an approach for entity disambiguation by strengthening low power identity criteria using event-based association-evidence (context) over ILNs.

This idea is developed in Section 3 and evaluated in Section 4 using entities of type person in historical datasets on ILNs with low discriminating identity criteria power. We conclude in section 5.

2 RELATED WORK

The use of contextual information to address the challenge of assessing the quality of links discovered has also been investigated by [10]. It uses a hybrid similarity measure that combines content-based and context-based similarities, the latter computed using *steady state probability of a random walk with restart*. The key point that sets our work apart from this approach is that we consider contextual information as an *event-based association-evidence* to *automate the removal of unsubstantiated links from a given network*. This usage of context is somewhat shared by the extensive work by Dong in [2]. Only, in our approach, *we stick with a single type of entity, the association-evidence is not restricted to a directly connected pair of references and we do not perform propagation.*

The work of Raad et al. in [9] on detecting erroneous identity links is *comparable*. It relies on link asymmetry (different data providers assert either the identity A→B or A←B) and on the density of the network as evidence for link quality. This is not applicable in our case as we assume that each identity link is symmetric.

Identity link networks can be created by sophisticated clustering algorithms such as CLIP[12]. However, the main point of this paper is not to generate the best possible clusters which yet needs to be evaluated, but rather to isolate sub-clusters in already computed clusters (ILNs) with a better chance of being true positives. In particular, CLIP is not applicable in our case as it does not allow for duplicated resources although duplication is a very common phenomenon in real-data, especially in non curated datasets.

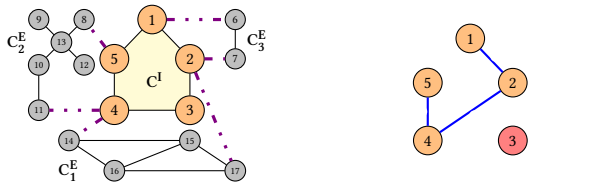
3 DYNAMIC ILNs ADJUSTMENT

A pair of nodes in an ILN C is said to be supported by association-evidence when event-based association relationships or mappings can be observed between C and one or more disconnected ILN graphs. Such a *mapping* involves two vertices, one from each of the ILN graphs. Depending on the viewpoint, one of the two ILNs takes the role of the *investigated ILN* C^I while the second one takes the role of the *association-evidence ILN* C^E. In a setting where such mappings occur, the mapped nodes under investigation are said to be *reconciled* based on the supporting mapped evidence nodes. This setting, composed of (1) an evidence set of vertices, (2) their respective links and (3) a pair of mapping relationships is what we define as the *event-based context* for corroborating an existing or virtual investigated link. *Reconciling a pair of investigated resources* then amounts to *corroborating the presence of the pair under a particular event-based context*. In the next two subsections we show the reconciliation based on a simple and then more complex context. Finally, we show in the last subsection how this idea of context and reconciliation is used to prune an ILN.

3.1 Context-based reconciliation

Figure 1a depicts four identity clusters, each composed of links generated under a weak identity criteria which does not provide enough ground to confirm that all nodes of the same cluster refer to the very same person. Let us assign the role of C^I to the cluster in the middle with orange nodes, and respectively the role of C_1^E, C_2^E and C_3^E to the clusters with gray nodes on the bottom, left and right of C^I. The purple dotted lines map nodes from C^I to their paired nodes in either of the C^E clusters indicating an event-association (e.g. marriage) between the matched nodes from the investigated cluster C^I and those from the evidence clusters C_1^E, C_2^E and C_3^E. The mappings between C^I and C_3^E depict the ideal reconciliation scenario, called a **ONE-TO-ONE EVIDENCE MAPPING**, where a pair of **DIRECTLY CONNECTED** nodes from C^I is validated by another pair of **DIRECTLY CONNECTED** nodes from C_3^E. In this scenario, where *additional event-based context is provided*, it is easy to conclude that the probability that the candidate nodes 1 and 2 in C^I are co-referent is strengthened given the context.

Making use of the event-based context becomes slightly complex when the mapping is **ONE-TO-MANY** rather than ONE-TO-ONE, as observed in Figure 1a with the mappings between C^I and C_1^E or C_2^E. For example, the nodes 4 and 5 from C^I are respectively mapped to the *indirectly connected evidence nodes* 11 and 8 from C_2^E via two intermediate evidence nodes, 10 and 13. In short, the idea here is to *reconcile* pairs of nodes ($n = 2$) using a connected set of nodes ($n \geq 1$) from an *evidence cluster*. This does not require the investigated pair to be directly connected as shown in Figure 1a.



(a) Extending by association an investigated cluster of candidate references selected by matching algorithm(s).

(b) Computing evidence strength for corroborated nodes.

Figure 1: From reconciliation strength to inferred reconciliation strength.

DEFINITION 1. *Let C^I and C^E be two undirected and connected candidate identity graphs, where C^I represents the cluster of candidate nodes under investigation while C^E is the cluster of association-evidence. A pair of nodes $(n_1, n_2) \in C^I$ is said to be corroborated by a sequence of connected nodes $(m_1, m_2, ..., m_i) \in C^E$ when the start and end nodes (m_1, m_i) are mapped to (n_1, n_2) via a set of relationships R that form a **cycle** (see algorithm 1). When the shortest distance between m_1 and m_i equals 1, the mapping is said to be a one-to-one mapping, otherwise, it is said to be one-to-many.*

3.2 Pruning & Extractions

Now that we know how to reconcile candidate pairs of investigated nodes, the next goal is to isolate identity sub-cluster(s) with high quality from an initial candidate identity graph (**Scenario-A**, Figure 1a), by eliminating potentially false positive resources. For that, we simply extract the evidence-based sub-graphs (**Scenario-B**, Figure 1b).

Let us assume C^I, the investigated cluster (the orange nodes in Figure 1), is composed of five nodes connected in a ring topology. Suppose that three links $\{e_1, e_2, e_3\} \in C^I$ are supported by association-evidence obtained through three evidence clusters (clusters of gray nodes). As shown in **Scenario-A**, there is association-evidence for reconciling four candidate nodes via the following edges: $e_1 = (1, 2)$, $e_2 = (2, 4)$ and $e_3 = (4, 5)$. Direct links exist only between elements of the pairs (1, 2) and (4, 5), while no direct link

Algorithm 1: Algorithm for cycle detection.

```
input  : a set Ξ mapping each clustered node to its cluster; a set Λ of associations.
output : a set Ω mapping a pair of clusters to their pairs of nodes that are supported by
         association evidence from Λ
begin
    Ω, Ψ ← ∅, ∅                                   /* initialize result and temporary sets */
    for (n1, n2) ∈ Λ do                           /* for each assoc. in Λ, O(|Λ|) */
        if n1 ∈ Ξ and n2 ∈ Ξ then                 /* if both nodes are in the mapping Ξ */
            C1, C2 ← Ξ[n1], Ξ[n2]
            if C1 < C2 then
                key, value ← (C1, C2), ← (n1, n2)
            else if C1 > C2 then
                key, value ← (C2, C1), ← (n2, n1)
            if key ∈ Ψ then /* if key has once been added to Ψ then it has cycle */
                if key ∈ Ω then /* if key has already been added to the result Ω
                   */
                    Ω[key].add(value)
                else
                    Ω[key] ← Ψ[key];
                    Ω[key].add(value);
            else
                Ψ[key] ← [value]
    return Ω
```

exists between elements of the pair (2, 4). **Scenario-B** illustrates the context-based reconciled nodes. While the pair (1, 2) is corroborated by the *directly connected evidence* nodes {(6, 7)}, the pair (4, 5) is corroborated by an association-evidence path {(8, 13), (13, 10), (10, 11)}. The last reconciliation materializes a *virtual link* between nodes of the pair (2, 4) $\in C^I$.

This illustrates how non-corroborated existing links are pruned and new reconciled links are created, all this resulting in excluding potentially wrong candidate resources, like node 3 in this example.

4 EVALUATION

Inputs . The *Golden Agents project*[1] *aims to analyze data collected between the 15th and 18th centuries, i.e. the Dutch Golden Age.* Among the data sources are registries from the Amsterdam City Archives[2](SAA, StadsArchief Amsterdam) and Ecartico[3]. In particular, three of the data sources accounting for about 7 million names (hereby referred as SAA datasets) mention couples in events that occurred in Amsterdam: (i) **Marriage** mentions husband and wife to be, previous husband and wife; (ii) **Baptism** mentions father and mother; and (iii) **Burial** mentions deceased and related person (e.g. husband or wife of). These sources are used to investigate the disambiguation process. For that purpose, we apply two filters: (i) events registered between 1600 and 1650 and (ii) mentions that occur also in a rather smaller yet curated dataset called **Ecartico -Authoritative Dataset-** consisting of around 30K disambiguated names of mostly notable people, including those living in Amsterdam during the dutch Golden Age (version of September 2017). This results in 12.5K mentions to disambiguate.

The SAA datasets are matched among themselves and against Ecartico with a 0.85 threshold for approximate string matching on names, producing about 120K links. The links are clustered using a straightforward clustering algorithm [4] that produces **1295** clusters. From those, 50.81% are of size 3 or bigger while **44** are bigger than 30 and account for over 106K links. The latter are not evaluated since they are expected to be manually classified as BAD. This reduces the links to be validated to 13.4K. Overall, 95% of the clusters are of size ≤ 19. The input data is available online.[4]

Effect of low power identity criteria . The goal of this experiment is to show that a given set of ILNs generated under poor discriminating criteria is *likely to contain a significant number of bad clusters.* To test this hypothesis, we use as input a set of **1251** (1295 - 44) ILNs for which we believe the applied content-based identity criteria to be weak, i.e. not enough to correctly discriminate the entities (for example, clusters of very common names). Next, a team of three experts (one historian and two non-domain-experts but with good understanding of the data) classified the generated ILNs as **GOOD** or **BAD**. The historian's expertise was required for 213 difficult clusters. Indeed, **BAD** clusters account for 61.4% for size $s \in [3, 30]$ and 20.7% for size 2. From the available data, the experts also identified **1145** clusters as ground-truth positives. The gold standard is available online[5].

Substantiated vs. unsubstantiated ILNs. The statistic above shows the significant inclusion of noise by the matching strategy due to the weak identity information, and supports our motivation. We now investigate the data further to test *"how likely are clusters with supported association-evidence to be correct as opposed to those with no observed association-evidence?".*

Here, we use as association-evidence person-names that are mentioned as couples in each record. To examine the effect of evidence by association, the 1251 clusters generated in the previous experiment are split into 987 clusters **WITH CYCLES** (see Definition 1) and 264 clusters **WITHOUT CYCLES**. Under this dissociation, the results show that for the 73 clusters of size $s \in [3, 30]$ where **NO CYCLE** is observed, no **GOOD** cluster is found. Also, the majority (56.19%) of clusters **WITH CYCLES** is still bad, since there are still false positives among them. Moreover, for clusters of size 2, the numbers reveal a *correlation between the presence of association-evidence and the increase in the number of* ***GOOD*** *clusters, while in its absence, the number of* ***BAD*** *clusters increases instead.*

From the 1251 ILNs, 742 are true positives (1145 ILNs are expected based on the ground truth), which translates into $F_1 = 0.62$ (where $precision = \frac{742}{1251} = 0.59$ and $recall = \frac{742}{1145} = 0.65$). If we assume naively that clusters with cycles are likely to be correct, we have 987 ILNs with cycles from which 680 are true positives out of 1072 ground-truth positives. These numbers equate to $F_1 = 0.66$ (where $precision = \frac{680}{987} = 0.69$ and $recall = \frac{680}{1072} = 0.63$). This second step with zoomed in numbers *reveals indeed potential benefits in the use of context as extra means for estimating the quality of an ILN* and indicates that we are on a good path, which is tested in the next section.

Dynamically adjusted ILNs. Now, we take a look at *how the proposed approach performs* by examining the quality of clusters it generates. In the previous analysis, we observed an improved prediction of the naive cycle-based prediction over the overall original prediction with weak identity criteria for detecting whether references grouped together are a correct representation of a single real object. Using the contextual information, we are now able to extract subsets (cluster splitting) of an input cluster for which we

[1]Creative Industries and the Making of the Dutch Golden Age is funded by the NWO Large-investments program (https://www.goldenagents.org).
[2]https://archief.amsterdam
[3]www.vondel.humanities.uva.nl/ecartico
[4]https://github.com/alkoudouss/K-CAP-2019/tree/master/Data
[5]https://github.com/alkoudouss/K-CAP-2019/tree/master/Gold-Standard

Clusters	Description	Links	Ground Truth
1206 ∃ cycle	$s = 2$	1206	*Good* : 1018 / *Bad* : 188
642 ∃ cycle	$s \in [3, 30[$	8109	*Good* : 353 / *Bad* : 289
1848	$s \in [2, 30[$	9315	

Table 1: Evaluation of the proposed approach.

have evidence. This leads to an increase in the number of clusters, from 1295 to **1848**, and a decrease in the total number of links, from 13343 to 9315, given the exclusion of references for which no association-evidence is observed. These newly generated clusters are again manually evaluated.

Table 1 shows that, with clusters of size two, association-evidence confirm that the large majority of the matched pairs are **good**. This shows that our approach works well. Furthermore, if we exclude ILNs for which the nodes belong to the previously excluded 44 ILNs of size bigger than 30, we now have **1219** reconciled ILNs. From these, the reconciliation approach is able to correctly find 1040 out of the 1145 ground truth positives. This leaves us with $precision = \frac{1040}{1219} = 0.85$, $recall = \frac{1040}{1145} = 0.90$, and $F_1 = 0.88$.

Throughout the first experiment, we show with the ground-truth statistics that ILNs generated under low identity criteria are filled with false positives, leading to bad identity clusters. In the experiment two, by just separating corroborated clusters from non corroborated ones, the increase of F_1 score from 0.62 to 0.66 indicates potential in using event-based context. Finally, in this last experiment, where we make use of context to split clusters and prune nodes, the further increase from 0.66 to 0.88 indicates that the use of event-based context and its implementation help in generating more reliable sub-clusters.

Overall, we investigated cases in which context did not help. These are scenarios in which big clusters (very common names) end-up corroborating one-another. In these cases the human judges used extra information such as the order of the marriage, baptism and burial dates to identify different couples with the same common names (bad reconciliation). Note that such information are not used by the entity matching algorithm.

5 CONCLUSION

In this paper, we present an approach for supporting entity disambiguation by strengthening low power identity criteria (content) using event-based association-evidence (context) over identity link networks. This idea translates into two interconnected tasks. First, *find* pair of nodes, directly or indirectly connected, for which association-evidence exist (**cycles between identity clusters**). Second, *extract* the **substantiated identity sub-cluster(s)**. Incrementally, through an experiment conducted in three steps, we show that the quality of the generated identity clusters *improves with the strengthening of the initial identity criteria by relying on context.*

Regarding the quality of the results and benefits for the Golden Agents project, the developed method has shown to be promising for supporting the challenging task of disambiguating Amsterdamers during the Golden Age on a larger scale than previously possible, despite being tested in a reduced experiment. Even though there are still a few false positives and negatives, we believe that this rate is now acceptable given that (i) manual evaluation of the complete datasets is impossible and (ii) the content-based identity criteria used for the initial identity clusters can still be improved, thus discarding some bad links such as the ones corresponding to life-events after the death date.

In this paper, the notion of cycle is drawn using a single evidence cluster. As future work, we plan to investigate *how to extend this idea of strengthening the evidence based on the notion of cycle to more than a single evidence cluster.* For example, an author A_0 has published four articles: two papers are published with a co-author A_1 and the remaining two with A_2. By using the proposed approach we have evidence for splitting the original cluster of A_0 into 2 sub clusters with higher confidence instead of keeping them as one with low confidence. However, if A_1 and A_2 happen to have a paper together, we now have a second degree evidence for merging all instances of A_0. This idea also opens new horizons for using annotated evidence with strength or semantic for reasoning for example.

5.1 Acknowledgement

We kindly thank *Albert Merono* for his constructive comments and *Judith Brouwer* for her help in creating the ground-truth used in this article. This work was supported by the European Union's 7th Framework Programme under the project RISIS (GA no. 313082) and by the Investment Subsidy NWO Large 2015-2016 under the Golden Agents project (no. 175.010.2015.009).

REFERENCES

[1] Manel Achichi, Zohra Bellahsene, and Konstantin Todorov. 2017. Legato results for OAEI 2017.. In *OM@ ISWC*. 146–152.
[2] Xin Dong, Alon Halevy, and Jayant Madhavan. 2005. Reference reconciliation in complex information spaces. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 85–96.
[3] Daniel Faria, Booma S Balasubramani, Vivek R Shivaprabhu, Isabela Mott, Catia Pesquita, Francisco M Couto, and Isabel F Cruz. 2017. Results of AML in OAEI 2017.. In *OM@ ISWC*. 122–128.
[4] Al Idrissou, Frank van Harmelen, and Peter van den Besselaar. 2019. Network Metrics for Assessing the Quality of Entity Resolution Between Multiple Datasets (Invited submission as extension of [5], Under Revision, Open Access). *Semantic Web Journal* (2019). http://semantic-web-journal.org/content/network-metrics-assessing-quality-entity-resolution-between-multiple-datasets
[5] Al Koudous Idrissou, Frank van Harmelen, and Peter van den Besselaar. 2018. Network Metrics for Assessing the Quality of Entity Resolution Between Multiple Datasets. In *Knowledge Engineering and Knowledge Management*, Catherine Faron Zucker, Chiara Ghidini, Amedeo Napoli, and Yannick Toussaint (Eds.). Springer International Publishing, Cham, 147–162.
[6] Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, and Valerie Cross. 2017. LogMap family participation in the OAEI 2017. In *CEUR Workshop Proceedings*.
[7] Abderrahmane Khiat and Maximilian Mackeprang. 2017. I-Match and OntoIdea results for OAEI 2017.. In *OM@ ISWC*. 135–137.
[8] Xinze Lyu, Qingheng Zhang, Wei Hu, Zequn Sun, and Yuzhong Qu. 2017. njuLink: results for instance matching at OAEI 2017.. In *OM@ ISWC*. 158–165.
[9] Joe Raad, Wouter Beek, Frank Van Harmelen, Nathalie Pernelle, and Fatiha Saïs. 2018. Detecting Erroneous Identity Links on the Web using Network Metrics. In *ISWC*. Springer, 391–407.
[10] Hossein Rahmani, Bijan Ranjbar-Sahraei, Gerhard Weiss, and Karl Tuyls. 2016. Entity resolution in disjoint graphs: an application on genealogical data. *Intelligent Data Analysis* 20, 2 (2016), 455–475.
[11] Shu Rong, Xing Niu, Evan Wei Xiang, Haofen Wang, Qiang Yang, and Yong Yu. 2012. A machine learning approach for instance matching based on similarity metrics. In *International Semantic Web Conference*. Springer, 460–475.
[12] Alieh Saeedi, Eric Peukert, and Erhard Rahm. 2018. Using Link Features for Entity Clustering in Knowledge Graphs. In *ESWC*. Springer, 576–592.