Combining Node Embeddings with Domain Knowledge for Identity Resolution

J. Baas[a], M. M. Dastani[a] and A. J. Feelders[a]

[a] *Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, Netherlands*

Abstract

The application of powerful and popular machine learning methods on real life historical data generates sub-symbolic models of the data. Such models do not perform well when trained with insufficient (or no) ground truth. We argue that the performance of these models could be improved by incorporating domain-specific knowledge, and propose an approach to incorporate symbolic domain-specific knowledge in the sub-symbolic models of the data. We show with experimental results on real historical data that our approach improves performance.

1. Introduction

Unsupervised machine learning methods can be applied on real world historical data, for example, to create models of the data in which duplicates of entities are resolved. However, the generated models may be suboptimal due to a lack of (sufficient) ground truth. For example, they may produce errors in the identity resolution outcome, where pairs of entities are identified as identical while in reality they are not. We argue that domain-specific knowledge can be used to detect and correct such errors and propose an approach to incorporate domain-specific knowledge in the identity resolution algorithm. Examples of such domain-specific knowledge are 1) Two entities in the same civil registration cannot be identical, and 2) Two entities, one with birth date x, the other with marriage date y, if $x > y$ then the two entities cannot be identical.

Additionally, the domain-specific knowledge itself can also contain inherent uncertainty. An example of ambiguity in the data causing the knowledge to be uncertain is when a burial registry, in which two persons are mentioned (one is dead and the other is the witness), does not say which person has died. Therefore, when we construct a rule to capture this knowledge, we do not treat the conclusion of said rule as 100% certain.

Graphs and Networks in the Humanities, February 3–5, 2022, Amsterdam, The Netherlands
j.baas@uu.nl (J. Baas); m.m.dastani@uu.nl (M.M. Dastani); a.j.feelders@uu.nl (A.J. Feelders)
https://github.com/jurian (J. Baas)
0000-0001-8689-8824 (J. Baas); 0000-0002-4641-4087 (M.M. Dastani); 0000-0003-4525-1949 (A.J. Feelders)

2. Method

2.1. Sub-Symbolic Knowledge

We assume a set of entities that occur in a data set to be embedded in a Euclidean space. We also assume that the embedding is created in such a way that the cosine similarity between the embedded entities reflect their similarities in the original data. An example of how such an embedding can be created is explained in our previous work [1]. We now consider the set of weights $s_{ij} \in [-1,\ 1]$, for each possible pair of entities i and j that occur in the data set, as the knowledge encoded in the sub-symbolic model of the entities. In the rest of this paper we use the term sub-symbolic knowledge to refer to these weights.

2.2. Symbolic Domain-specific Knowledge

We recognize two categories of domain knowledge:

1. *Meta knowledge about the data set.* An example is when we know that (part of) a data set is already disambiguated. We can therefore conclude that any two entities originating from this set cannot be identical.

2. *Knowledge rules that govern the entities described in the data set.* For instance, one cannot occur in a record before they are born, or, as an active participant, occur after they have died. Note that the validity of such rules hinges on the interpretation of the meaning of a record. For example, when examining records about books, a long deceased person can be mentioned as the author of a book, and our previously mentioned knowledge rule does not apply.

Some knowledge rules can fit into both categories, depending on their interpretation. For example, consider the rule that two entities co-occurring in the same record cannot be identical. We can either regard this as a property of the records in the data set, or, for example, as a property of marriage. That is, one cannot marry oneself or declare oneself as deceased. We assume rules to have the form $r(i, j) \rightarrow i \neq j$, where r is a relation between entities i and j. For example, $r(i, j)$ may represent the fact that i and j occur in the same record. The reading of the rule is then *if i and j occur in the same record, then it can be ruled out that i and j are the same.*

2.2.1. Knowledge Rule Intuitions

We describe two kinds of intuition concerning the application of knowledge rules to the entities.

1. The first intuition suggests that the application of knowledge rule only impact precision, while recall remains constant. This is because errors are potentially caught by the rules and that novel links are never introduced by them.

2. The second intuition suggests that precision would either remain constant or increase. There are either no occurrences of entities where these rules apply or they catch some errors.

Both fortunately and unfortunately, both these intuitions are false. The first intuition does not take into account the interaction with the sub-symbolic knowledge. That is, ruling out some possible matches will make other previously disregarded matches as potential candidates. The resolution algorithm may exploit this effect and potentially increase recall in addition to precision. The second intuition disregards the interaction with errors in the data set and (correct) judgements of domain experts when they constructed the ground truth. Even when a rule is valid in a certain domain, it may introduce results that differ from the ground truth when the data is erroneous. For example, a match has been soundly ruled out and is absent from our result, meanwhile, a domain expert looks at the same data and concludes that the data was probably in error, and makes the match anyway. This causes precision to decrease, as the number of true positives has decreased while the number of false positives remained constant.

2.2.2. Uncertain Rules

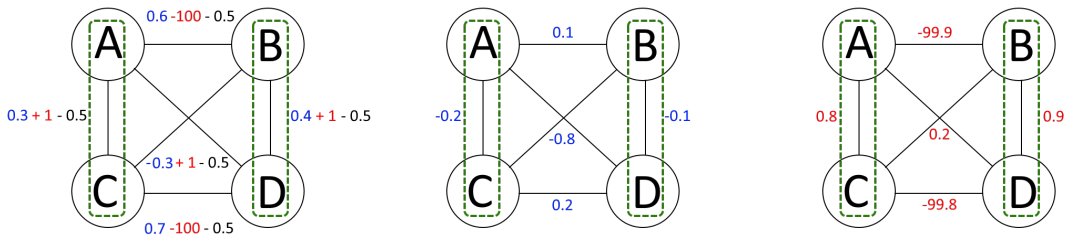
When further rules are considered that have a probability less than 1 of being correct if fired, an extra dimension of complexity is introduced along those mentioned above. Aside from their possible positive effects, both precision and recall can potentially decrease as errors go unnoticed or correct links are wrongfully ruled out. The results of applying all the rules on a potential link then have to be aggregated for a final verdict on the probability of correctness for that link. We use $\mathcal{R}_{ij} = \{r_1, \ldots, r_k\}$ to denote the set of all knowledge rules that are applicable to entities i and j. Moreover, we assume function p, defined as $p : \mathcal{R}_{ij} \rightarrow [0, 1]$, to denote *the probability of the conclusion of knowledge rules*. So, $p(r)$ is *the probability of the conclusion of the knowledge rule with* r. Since the conclusion of all rules are the same, $p(r)$ denotes the probability of ruling out that i and j are the same. For example, $p(r) = 0.5$ should be read as in 50% of the cases, the conclusion of rule r is not correct. When evaluating a set of rules on a potential pairs of entities, denoted by the pair of entities i and j, the outcomes are aggregated with the following function:

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

where, p_{ij} is the total probability that we can rule out that entities i and j are the same and we assume independence between observations.

2.2.3. Supreme Rules

Lastly, we introduce the concept of *supremacy of rules*. A rule is supreme if it is known beforehand that the rule cannot possibly be wrong under any circumstance, and that it should override (and not be combined with on equal terms) any sub-symbolic knowledge



(a) Constituent weights (b) Only sub-symbolic weights (c) Combined weights

Figure 1: An example of how symbolic (red) and sub-symbolic (blue) weights are combined with the cutoff (black) value. The ground truth clusters are denoted with the dashed green lines and both diagonals have the same weight. Without the (red) symbolic weights, panel b, the optimal clustering is (AB) and (CD). However, when including the symbolic weights, panel c, the optimal clustering becomes (AC) and (BD).

that is in conflict with it. The notion of a supreme rule is relative to a data set and domain dependent. An example of a supreme rule is that, if it known beforehand that part of a data set is perfectly disambiguated, any two entities originating from that part cannot possibly be the same entity, regardless of what sub-symbolic knowledge says. We combine supreme rules with sub-symbolic knowledge by assigning a very large negative weight to potential links on which a supreme rule has fired, such that eliminated links are never part of the optimal solution in a clustering algorithm.

2.3. Combining Symbolic and Sub-symbolic Knowledge

With the above two points we can construct an equation for calculating the associated weight w_{ij} for a given pair of entities i and j:

$$w_{ij} = \begin{cases} -1e^6 & \text{if any supreme rule fired for pair i, j} \\ \tanh\left(\log \frac{p_{ij}}{1-p_{ij}}\right) & \text{otherwise} \end{cases} \tag{2}$$

Equation 2 yields a very large negative weight if any supreme rule fired for the pair of entities i and j. In all other cases, it transitions smoothly between $w_{ij} = 1$ for $p = 1$, and $w_{ij} = -1$ for $p = 0$. Lastly we combine the symbolic weight w_{ij} with the sub-symbolic weight s_{ij} by adding them together and then subtracting a cutoff value which allows us to control the size of clusters. Note that, if $p = \frac{1}{2}$, then $w_{ij} = 0$ and only sub-symbolic knowledge is used. The intuition behind this is that when uncertainty about symbolic knowledge is at its maximum, its contribution is zero. Figure 1 shows an example how a small cluster is split up by a clustering algorithm. Note that in panel 1c the inclusion of domain knowledge has increased both precision and recall.

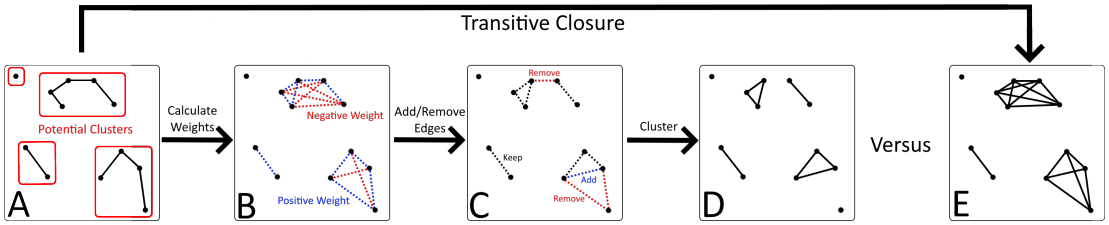


Figure 2: An overview of how potential clusters are potentially split up. We start in panel A, where we either, as a baseline, consider each potential cluster as a cluster without modification (panel E), or we generate weights (both symbolic and sub-symbolic) inside each potential cluster (panel B). These weights are then used by a clustering algorithm to add and remove links (panel C). The result in panel D is then compared to the baseline.

3. Experimental Setup

3.1. Data

For our experiments we have used four data sets containing real historical data from the cultural heritage domain. The City Archives of Amsterdam[1] is a collection of registers, acts and books from the 16^{th} century up to modern times. We made use of a subset collected from three different registers: Burial, Prenuptial Marriage and Baptism covering a period in the 17^{th} century. The burial register does not describe who was buried where, but records the act of someone declaring a deceased person. To this end, it mentions the date and place of declaration and references two persons, one of whom is dead. Sadly, it does not tell us which one of the two has died. The prenuptial marriage records tell us the church, religion and names of those who are planning to get married. It also mentions names of persons involved in previous marriages if applicable. The baptism register mentions the names of the parents and date of the event. Lastly the above records were combined with a subset from the Ecartico [2] data set containing marriage records. Ecartico is a comprehensive collection of biographical data from more well-known people from the Dutch golden age. For our experiments we use a subset containing 12,517 entities referring to (non-unique) persons, and a partial ground truth of 1073 clusters, obtained with manual validation by domain experts [2].

3.2. Potential Clusters

As mentioned before, we use a cutoff value to influence the size of the final clusters. This cutoff value also determines the size of the input potential clusters. That is, given a cutoff value, we start with a corresponding set of potential clusters, as shown in panel A of figure 2.

[1] https://archief.amsterdam/indexen
[2] http://www.vondel.humanities.uva.nl/ecartico

Table 1
All the rules used in our experiments

Entity 1	Comparator	Entity 2	Probability	Supreme
in Ecartico	and	in Ecartico	NA	yes
record	equals	record	NA	yes
marriage date	later than	death date	1	no
birth date	later than	death date	1	no
prenuptial marriage date	later than	death date	1	no
baptism date	later than	death date	1	no
prenuptial marriage date	later than	burial date	0.5	no
marriage date	later than	burial date	0.5	no
baptism date	later than	burial date	0.5	no
burial date	earlier than	death date	0.5	no
baptism date	earlier than	birth date	1	no
prenuptial marriage date	earlier than	birth date	1	no
marriage date	earlier than	birth date	1	no
burial date	earlier than	birth date	1	no

3.3. Rules

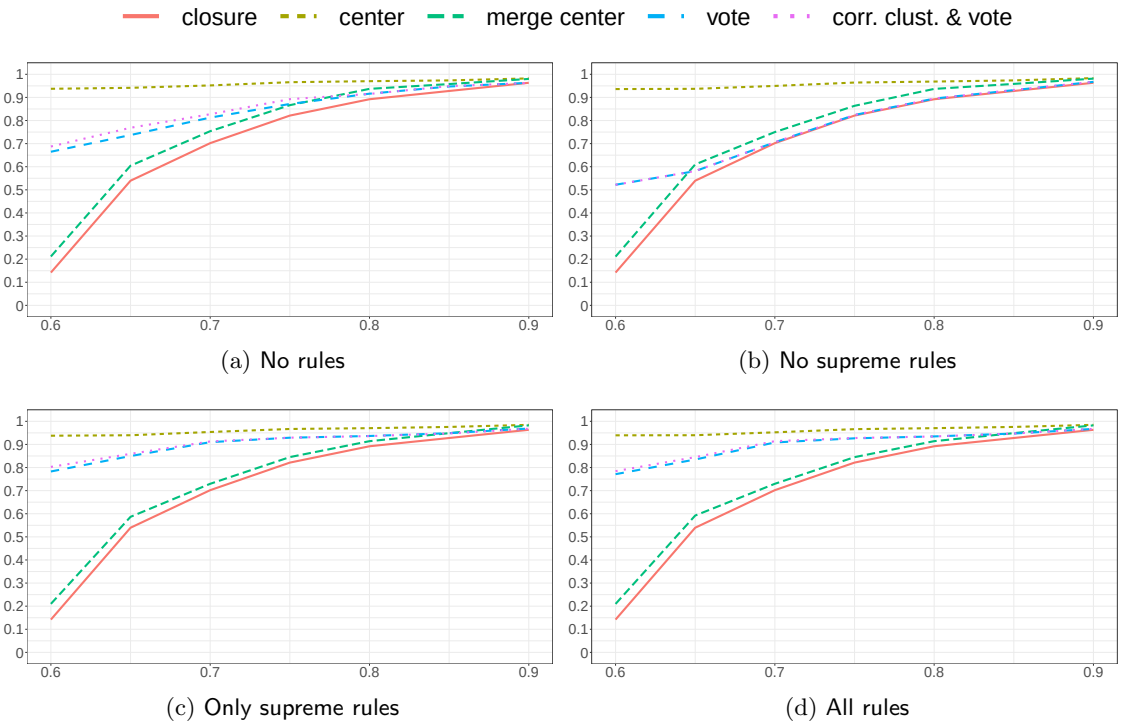
Table 1 shows all the rules used in our experiments. We have designed two supreme rules, namely that two entities co-occurring in the same record are not the same, and that two entities originating from Ecartico are not the same. Furthermore, there are 12 rules that consider the dates events took place in, and whether one occurs after the other. The rules considering burial records are not 100% certain and we have given them a probability of $\frac{1}{2}$. One exception is comparing dates in burial records to birth dates, as neither the deceased person nor the person who declared the death can perform their respective roles before they are born. We perform four experiments, each with a different set of rules: a) no rules, b) only non-supreme rules, c) only supreme rules, and d) all rules.

3.4. Clustering Algorithms

We use four different clustering algorithms to split up potential clusters. Each will take a potential cluster as input, and then, using the internal weights, output one or more clusters. We go into each in more detail in our previous work [1].

4. Results

Figures 3 and figure 4 show the results of our experiments. The solid red line is the baseline performance, i.e. the transitive closure. We will discuss each figure in more detail below.



(a) No rules (b) No supreme rules

(c) Only supreme rules (d) All rules

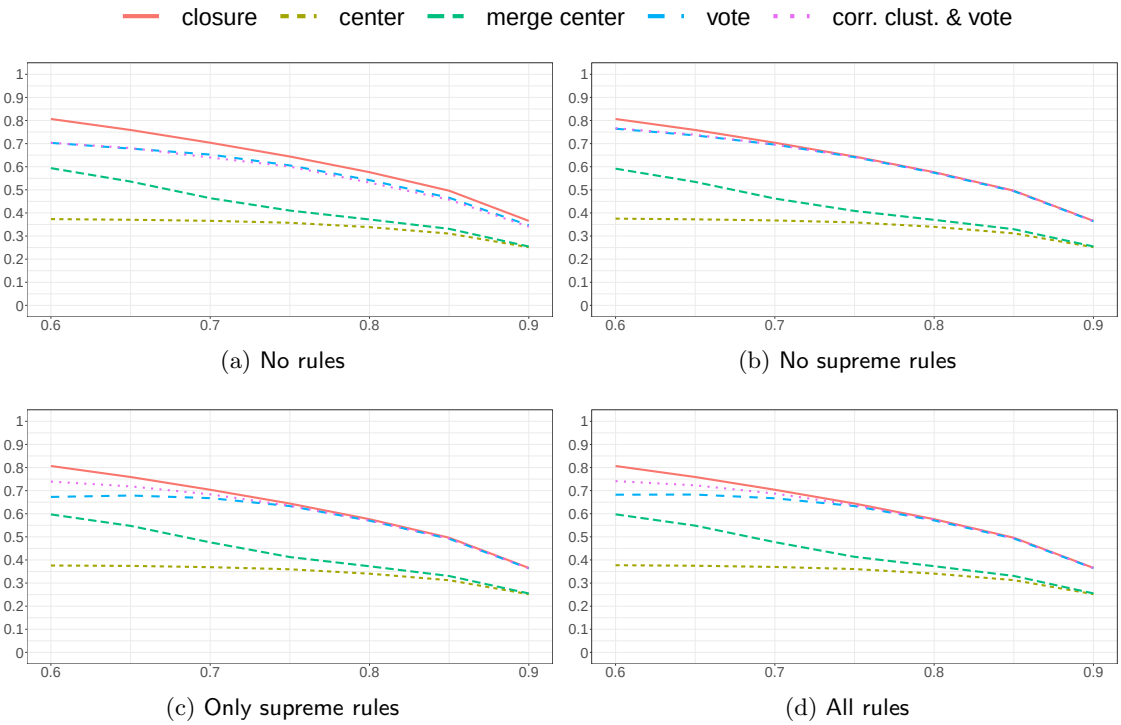
Figure 3: Precision (y-axis) for a range of cutoff values (x-axis)

4.1. Precision

Figure 3 shows, for a range of cutoff values, how the precision changes for each clustering algorithm. Note that the supreme rules have a very large effect on precision, while the other rules have a detrimental effect over using no rules at all. From a manual examination we have determined that this is most likely due to these rules being in conflict with errors in the data. That is, the domain expert creating the ground truth observed that the dates in two records were not strictly correct for making a match, but, based on other information, concluded that this was indeed a match.

4.2. Recall

Meanwhile, figure 4 shows, for a range of cutoff values, how the recall changes for each clustering algorithm. Note again that the supreme rules can have a positive effect on recall when using the correlation clustering algorithm. As laid out in figure 1. The non-supreme rules have a positive impact on recall for both the vote and correlation clustering algorithms. However, combining both types of rules seems to diminish these outcomes.



(a) No rules (b) No supreme rules

(c) Only supreme rules (d) All rules

Figure 4: Recall (y-axis) for a range of cutoff values (x-axis)

5. Conclusions

We have shown that it is possible to combine symbolic and sub-symbolic knowledge in such a way to improve performance. 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 to extend the non-supreme rules to allow for more lenient comparison, such as "*x has to be at least 3 years after y for concluding they are not the same*".

References

[1] J. Baas, M. M. Dastani, A. J. Feelders, Entity matching in digital humanities knowledge graphs, Proceedings http://ceur-ws. org ISSN 1613 (2021) 0073.

[2] C. Latronico, V. Zamborlini, A. Idrissou, Amsterdamers: from the golden age to the information age via lenticular lenses, in: Digital Humanities Benelux, 2018.