A Method for Record Linkage with Sparse Historical Data

Giovanni Colavizza, Yannick Rochat, Maud Ehrmann

[1]DHLAB – Ecole Polytechnique Fédérale de Lausanne (EPFL)
CDH, INN, Station 14, CH-2015 Lausanne, Switzerland

{name.surname,name.surname}@epfl.ch

1. Introduction

Massive digitization of archival material, coupled with automatic document processing techniques and data visualisation tools offers great opportunities for reconstructing and exploring the past. Unprecedented wealth of historical data (e.g. names of persons, places, transaction records) can indeed be gathered through the transcription and annotation of digitized documents and thereby foster large-scale studies of past societies. Questions such as how and where people lived, what were their occupations, how were they linked to each other, etc. can now be addressed [adverb]. Yet, the transformation of hand-written documents into well-represented, structured and connected data is not straightforward and requires several processing steps. In this regard, a key issue is entity record linkage, a process aiming at linking different mentions in texts which refer to the same entity. Also known as entity disambiguation, record linkage is essential in that it allows to identify genuine individuals, to aggregate multi-source information about single entities and to reconstruct networks across documents and document series.

In this paper we present an approach to automatically identify coreferential entity mentions of type *Person* in a data set derived from Venetian apprenticeship contracts from the early modern period (16th-18th c.). Taking advantage of a manually annotated subpart of the document series, we compute distances between pairs of mentions, combining various similarity measures based on (sparse) context information and person attributes. Section 2 motivates our work, section 3 presents the data-set, section 4 describes the method, section 5 presents the results and section 6 concludes and considers future work.

2. Historical entity record linkage/Task definition and objective

Major challenges when dealing with people-related data are homographic person names referring to different persons, as well as the existence of name variants referring to the same person. These are well-known issues in the field of Natural Language Processing for which various approaches have been devised, first via mention clustering [**?**, **?**], more recently via linking to a knowledge base [**?**, **?**].

In the context of historical data, dealing with person name ambiguity is all the more difficult since data is inherently sparse and uncertain (resulting in poor mention context) and since knowledge bases such as DBpedia [**?**] contain very little about past average laypersons (resulting in poor entity context). It is however an essential step prior to any historical data analysis [**?**], which we address as part of the *Garzoni* project. This project aims at studying apprenticeship in early modern Venice by extracting information from archival material. Part of this material have been manually annotated[1], including

[1] The manual annotation phase is on-going.

count	whole period	1586-1600
# annotated contracts	11,525	2,687
# mentions	31,952	7,589
# entities	26,641	6,599
# entities with # mention $>$ 1	-	-
AVG mention per entity	1.09	1.08
AVG mention per entity with # mention $>$ 1	2.44	2.38
# uniq pair of mentions	861	382

Table 1. Entity-Mention stastistical profile for the whole vs. selected period.

mention links towards unique entities. Starting from a subset of the current data, we present a method for person record linkage, with the objective to complement its disambiguation coverage and to bootstrap a system to better automate entity disambiguation during annotation, in a active learning fashion.

3. The *Accordi dei Garzoni*

The *Accordi dei Garzoni* is a document series from the State Archives of Venice which originates from the activity of the *Giustizia Vecchia* magistracy. This judicial authority was in charge of registering apprenticeship contracts in order to protect young people while they were trained and/or providing domestic services [**?**]. As a result of this regulation, information for much of apprenticeship arrangements got centralized, today reflected in a exceptionally dense and complete archival series.

The *Accordi* consists of about 55,000 contracts registered from the year 1575 until 1772. Each contract features an apprentice, a master and often a guarantor, sometimes two. For the most part, these written records follow the same pattern and report on various attributes about persons mentioned (name, age, geographical origins, profession, etc.) and about contract terms (date, duration, financial conditions, etc.)

As of today, ca. 11,000 contracts have been manually annotated[2] and the resulting data is stored in a RDF triple store. What interests us here regards the annotation strategy for person entities: for each person mentioned in a contract, annotators create a *person mention* and, importantly, have to link it to a *person entity*. They do so either by selecting an already existing entity in the database or by creating a new one. Disambiguation towards an existing entity is supported by a name-based autocompletion mechanism; although very helpful, it does not handle name variants and annotators sometimes have to figure themselves which name to search for. The annotated dataset can therefore be considered as correct but not as exhaustive. The present work considers annotated documents from the period 1586-1600 (where many contracts have been annotated), for which statistics about contracts and entity/mention ratio are shown in Table 1. We used this data-set as our golden-set for our experiments.

[2] initially via a Semantic Media Wiki, now via a dedicated transcription and annotation web interface.

Feature	Variable type
firstname	string
surname	string
patronymic	string
gender	categorical
age	integer
profession	categorical
geographical origins	string

Table 2. Mention-level features

Feature	Variable type
workshop toponym	string
workshop parish	string
workshop sestriere[3]	string
workshop insigna	string
contract year	integer
contract duration	string
master profession	categorical

Table 3. Contract-level features

4. Approach

Given a set of mentions, our objective is to estimate the likelihood that two mentions refer to the same entity. We represent each mention by a vector of features and compare them pairwise using various similarity measures in different settings. The list of selected features at mention and contract levels are presented in Table 2 and Table 3 respectively.

With respect to our data-set and features, several point should be emphasized. First, data sparsity: it is common for a mention to have just a few features. Then, features are not evenly sparse (cf. Figure 1) and do not contribute equally to a possible linkage. Core features such as *name*, *surname*, *patronimic*, *gender* and *profession* should strongly correspond to consider a link as reliable. On the other hand, rare features such as *workshop insigna* can be very informative when shared by two mentions and should also be valued by the linkage algorithm. Finally, features are dependent, particularly on the role of the person (e.g. age indicated only for apprentices).

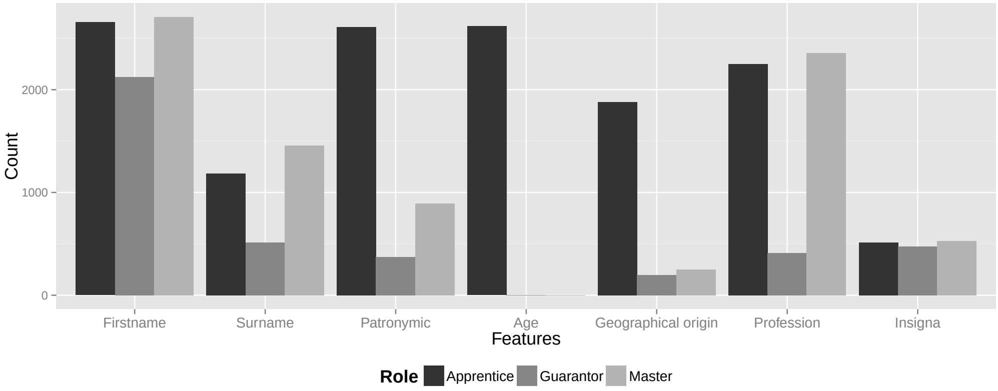

Figure 1. Distribution of features by role.

We construct three matrices of size NxN, where N is the number of mentions in the dataset. The first matrix Φ, the *feature matrix*, stores similarity scores of mentions pairwise. Scores are computed using measures over features, as follows:

- *year of contract*: the feature-score is measured via distance and diminishing returns. Each year of distance between 1 and 15 and between 15 and 30 decreases an initial score of 1 by 0.01 and 0.025 respectively, with a definitive cut-off after 30. For example, two contracts from 1590 and 1594 have a score of 0.096.

- *age*: similarly as per year, each year of distance of the difference between two ages decreases an initial score of 1 by 0.1.
- *gender* and *profession*: the feature-scores of these categorical feature are based on exact matches.
- *name*, *surname*, *patronymic* and *workshop toponym*: the feature-score is based on the Deverau-Levenshtein string-to-string distance measure.
- *geographical origins* and *insigna*: the feature-score is based on the Jaro-Winkler string-to-string distance measure.

The resulting feature-score vector of each pair is then normalized using the L_2 norm, whose value is taken as the score of the pair stored in Φ.

The second matrix Γ, the *combination matrix*, stores values that indicate whether a pair of mentions shares similar feature combinations or not. To build such matrix, we leverage the golden-set and identify combinations of features which produced a linkage on a role-by-role basis (e.g. master-master or guarantor-master). Features are considered activated when their feature-score is equal or above 0.84[4] and we filter out combinations occurring once. The score of a mention pair in Γ is 1.0 if it corresponds to valid combination of activated features for the given role pair; 0.5 if the role pair does not match but the combination is a valid one; 0.0 otherwise. This matrix accounts for feature dependencies and the different ways to name a person with respect to his/her role.

The third matrix Δ, the *filtering matrix*, scores mention pairs according to the number of activated core features (1.0 if $3+$ features (out of 5) are activated, 0.0 ortherwise[5]).

Given the three matrices (individually normalized), we consider the following function to compute the similarity score of a mention pair p:

$$S(p) = \delta_p[\lambda\pi_p + (1-\lambda)\gamma_p]$$

where δ_p is a boolean parameter taken from Δ which activates the filter over core features for pair p; π_p is the feature score taken from Φ; γ_p is the combination score from Γ; and λ is a parameter giving priority over vector features or combinations of features. $\delta \in \{0,1\}$ and $0 \leq \lambda$,π,$\gamma \leq 1$. This function allows us to adjust the different parameters: core vs sparse features (δ), feature norm (π) and feature combinations (λ).

5. Experiments

5.1. Evaluation

We evaluate our approach in terms of coverage and precision. With respect to coverage, we verify our method over 100 thresholds from 0.99 to 0.0. For each threshold, we compare linkage curves as the percentage of links obtained over the total possible with the coverage of the golden set. Precision is computed based on manual annotation of 50 randomly selected linkages.

[4] It has been shown in comparable setting that edit distance with cut-off at distance 3 (which for us is distance $\geq$ 0.85) provides good results [**?**].

[5] Features are activated when theirsimilarity is above 0.84.

	δ active			δ not active		
	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
all	0.21	0.3	0.21	0.0	0.26	0.15
w-o A-A	0.22	**0.61**	0.22	0.0	0.48	0.67*

Table 4. Precision with threshold ≥ 0.9 (* means not-significant statistics).

Both procedures are repeated with $\lambda \in \{0.1, 0.5, 0.9\}$ and δ activated or not, for a total of 6 configurations. The objective is to understand the individual contributions of the three components (core-features, norm features and combination features) to our function.

5.2. Results and Discussion

Results for the first and second evaluation procedure are presented Figure 2 and Table 4 (resp.). The best precision (0.61 and 0.48 in Table 4) is obtained with a balance of feature combinations and norm weights ($\lambda = 0.5$). δ proves very useful for filtering the input space (from 28.792.666 possible pairs to 44.263), and lowers the number of false positives, especially for links between apprentices (due to the strength of sparse features). The combination of the two (filtered input space and equal weights) provides the best results, especially for masters and guarantors. Linkage curves can be explained similarly: low λ entails a step-like curve (three steps at 0.0, 0.5 and 1.0), while high λ creates a Gaussian over the disambiguation space.

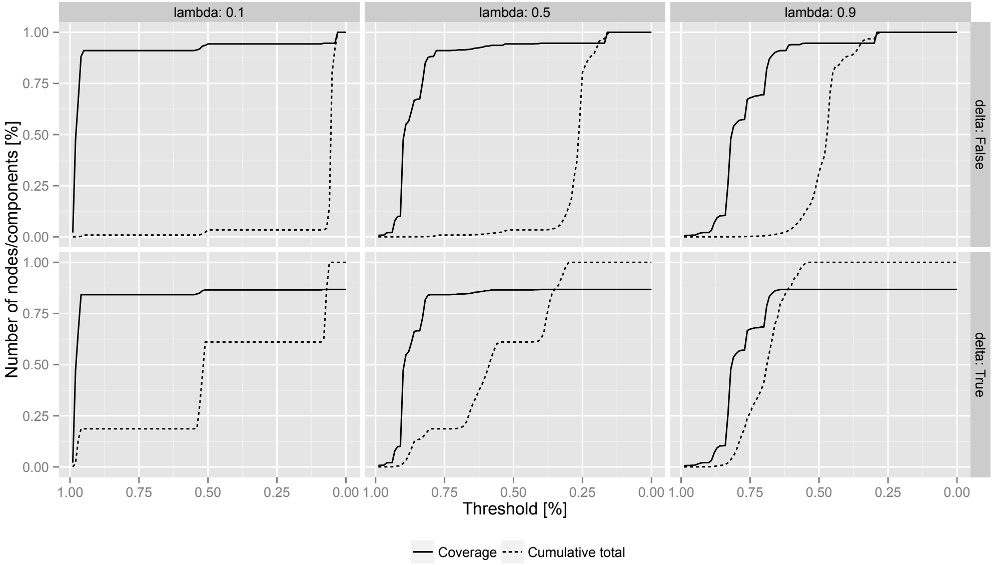

Figure 2. Linkage curves for the 6 parameter settings, over thresholds.

The results confirm that a balanced approach might be the best solution in a setting where data is sparse (high λ), the golden set is present but of limited coverage (low λ), and some prior assumptions on the required features can be made (δ). As shown in Figure

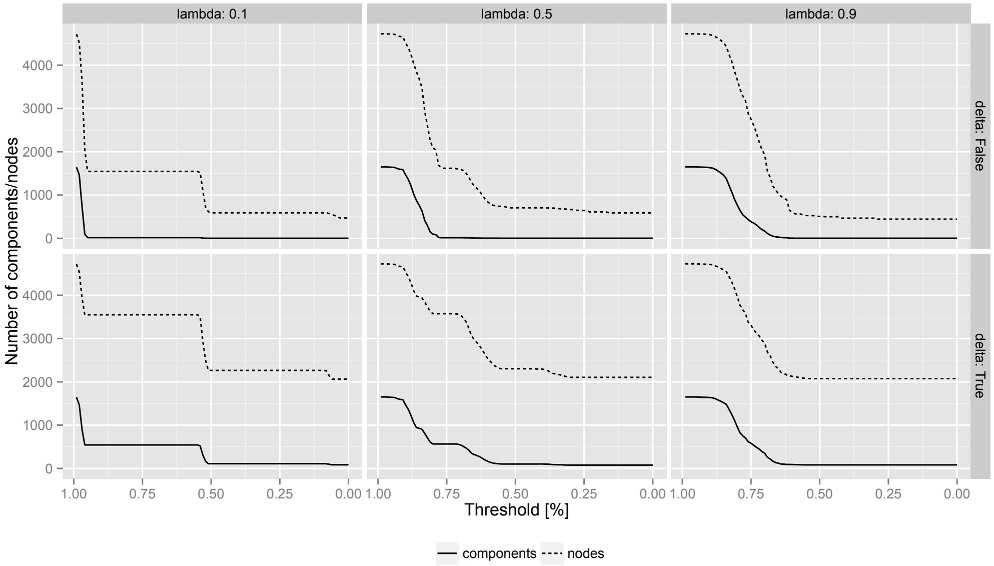

Figure 3. Graph properties for the 6 parameter settings, over thresholds.

3, the graphs with $\lambda = 0.5$ and $\delta = True$ collapse more gradually, providing the widest effective linkage space to explore. Eventually, results also suggest to proceed in an active learning fashion, where the system learns iteratively with new data as part of the golden set.

Finally, in order to further motivate our work, Figure 4 shows the largest components of the deduced social network with and without automatic disambiguation.The linkage method has the nice property of enlarging small components instead of the large ones.

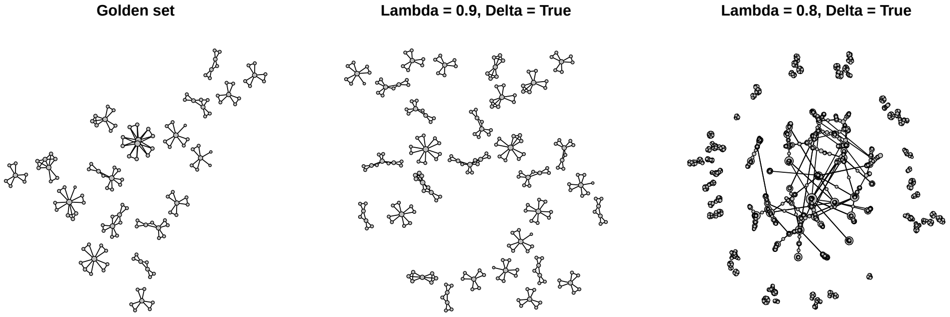

Figure 4. Largest components of social networks from golden-set (left-most) and from disambiguated data-sets (center and right-most).

6. Conclusion and Future Work

This paper presented a method to perform record linkage over mentions of persons from sparse historical data. The proposed system scales and is able to deal with different constraints such as data sparsity, variable prior knowledge and non-exhaustive golden-set. As future work we plan to apply the system to the whole dataset and to integrate it into the transcription and annotation interface, in order to use it for live, aided record linkage.

References

Artiles, J., Sekine, S., and Gonzalo, J. (2008). Web people search: results of the first evaluation and the plan for the second. In *Proceedings of the 17th international conference on World Wide Web*, pages 1071–1072. ACM.

Bellavitis, A. (2006). Apprentissages masculins, apprentissages féminins à venise au xvie siècle. *Histoire Urbaine*, pages 49–73).

Ji, H. and Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1148–1158. Association for Computational Linguistics.

Kleanthi, G., van der Burgh, B., Meeng, M., and Knobbe, A. (2015). Record linkage in medieval and early modern text. In *Population Reconstruction*, pages 173–195. Springer.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2013). Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

Mann, G. S. and Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proceedings of the* 7^{th} *Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 33–40.

Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460.