A Baseline Method for Genealogical Entity Resolution

Julia Efremova[1], **Bijan Ranjbar-Sahraei**[2], **Frans A. Oliehoek**[2], **Toon Calders**[1,3], **Karl Tuyls**[2,4]
[1]Eindhoven University of Technology, The Netherlands
[2]Maastricht University, The Netherlands
[3]Université Libre de Bruxelles, Belgium
[4]University of Liverpool, United Kingdom
i.efremova@tue.nl, {b.ranjbarsahraei,frans.oliehoek}@maastrichtuniversity.nl,
toon.calders@ulb.ac.be, k.tuyls@liverpool.ac.uk

Abstract

In this paper we study the application of entity resolution (ER) techniques on a real-world multi-source genealogical dataset. Our goal is to identify all persons involved in various notary acts and link them to their birth, marriage and death certificates. In order to evaluate the performance of a baseline approach based on existing techniques, an interactive interface is developed for getting feedback from human experts in the field of genealogy. We perform an empirical evaluation in terms of precision, recall and F-score. We show that the baseline approach is not sufficient for our purposes and discuss future improvements.

1 Introduction

Entity Resolution (Getoor and Machanavajjhala, 2012; Bloothooft, 1995; Christen, 2012) is the process of connecting disparate data sources for understanding possible identity matches and non-obvious relationships. Entity resolution is also known as duplicate detection (Christen, 2012) when discovering records that refer to the same entity occurs within a single database. ER is an important research problem in data mining and machine learning communities in the last decades

ER has been used in many different applications such as data warehousing, business intelligence, digital libraries, medical research and social networks. Recently, ER has found its way into genealogical data research as well (Ivie et al., 2007): In historical documents the real person entity could be mentioned many times, for instance in civil certificates such as birth, marriage and death certificates or in property transfer records and tax declarations. Usually, no common entity identifiers are available, therefore the real entities have to be identified based on alternative information (e.g., name, place and date). Considering the fact that the information is noisy, in large volumes of data the identification of duplication is difficult.

The past decade has seen some increased interests in Genealogical ER. Sweet et al. (2007) used an enhanced graph, based on genealogical record linkage, in order to decrease the amount of human effort in data enrichment. Moreover Schraagen et al. (2011) predicted record linkage potential in a family reconstruction graph by using the graph topology. Lawson (2006) used a Probabilistic Record Linkage approach for improving performance of information retrieval in genealogical research.

However, the mentioned works in Genealogical ER have mostly focused on linking references with *homogeneous structures:* all records have the same number of attributes which can be used for comparing two references. In this paper, in contrast, we are interested in applying ER to a real-world dataset with a *heterogeneous structure:* different references come from qualitatively different sources and references no longer have a similar data structure or identical sets of attributes. We refer to this problem as multi-source ER.

In particular, we are interested in performing multi-source ER on a database of historical records of a Dutch province called Noord-Brabant. As an example consider a person named *Theodor Werners* born in *Erp* on *August 11th, 1861*. He got married to *Maria van der Hagen* in *1888*. *Maria Eugenia Johanna Werners* was their child, born in *Erp* in *October 1894*. Two years after child's birth, they bought a house in *Breda*. *Theodor* died in *Breda* on *September 1st, 1926*. All information present the corpus is distributed over different sources such as civil certificates and notary acts. Applying ER to such a problem faces many challenges such as name alternatives, misspellings, missing data and redundant information.

The former have a structured form, but the latter consist mostly of free text, and as such are qualitatively different. For a given set of notary acts, we aim at identifying all persons involved and link them to their birth, marriage and death certificates.

The goal of this paper is to investigate in how far the above real-world multi-source ER task can be addressed with standard techniques. To that end, we propose what we refer to as a baseline method: a combination of standard techniques for the different phases of the ER process. We evaluate how well this baseline method performs on our real-world multi-source ER task. We conclude that is performance is far below acceptable and that, therefore, further improvements to multi-source ER techniques are needed, some of which we discuss for future work.

The remainder of this paper is structured as follows. In Section 2 we begin by describing the motivation of a real-life ER application. Then we describe standard ER techniques in Section 3 and demonstrate implementation of our baseline approach. In Section 4 we describe experiments and introduce the tools developed for historians to label data. In Section 5 we present evaluation of results. Section 6 offers a discussion about drawbacks and potential extensions of the proposed baseline-approach. Concluding remarks are included in Section 7.

2 Motivation: A Real-Life ER Application

The genealogical data which is used in this paper is provided by Brabants Historisch Informatie Centrum (BHIC)[1]. The data consists of two main different sources. The first source, civil certificates, is comprised of the birth, marriage and death certificates belonging to North Brabant, a province of the Netherlands, in the period 1700 - 1920 (in total around 1,900,000 certificates which provides 7,500,000 person references). The detailed information mentioned in each document varies very much. The different types of certificates contain different kinds and different amounts of information (structural differences). For example, in death certificates can be mentioned only the name of a deceased person on a specific date / in a specific place with unknown mother and father. In contrast some marriage certificates include many details such as groom's name, groom's age, groom's profession, the place and date of his birth and his parents details, and the same details for the bride.

Additionally, certificates of each type can contain empty or inaccurate values (data quality), for instance, person name or place can be misspelled.

A sample civil certificate is shown in Table 1.

Table 1: An example of civil certificate showing the birth data.

Person Name	Teodoor Werners
Gender	son of
Place of Birth	Erp
Date of Birth	14-04-1861
Father Name	Peter Werners
Father Profession	shopkeeper
Mother Name	Anna Meij
Mother Profession	-
Certificate ID	6453
Certificate Place	Erp
Certificate Date	16-04-1861

Structural differences and data quality problem in civil certificates influence the matching strategy and accuracy in different ways and both affect final results.

Another source of available information is a dataset of notary acts, which consists of around 90,000 free-text documents of the North Brabant province before 1920. These free-text documents include information about people involved in property transfers, loans, wills and etc. Since these documents are in a free-text format, all details are mentioned implicitly. An example of a notary act is shown in Table 2.

Table 2: An example of a notary act.

Theodor Werners, burgemeester van Boekel en Erp, wonend te Boekel bekent schuldig te zijn aan gemeente Erp Fl. 200,–. Waarborg: woonhuis, tuin, erf, bouw- en weiland Dinther en bouw- wei- en hooiland te Boekel	
TextID	100
Place	Boekel
Date	24-07-1896

The task we consider is to automatically find, for all persons mentioned in each notary act, the references in the civil certifications that correspond to the same persons. That is, we aim to find for every person its birth, death and marriage certificates, as well as those certificates where such a person is mentioned in a different role (e.g., father of a bride, etc.)

[1] http://www.bhic.nl/

3 A Baseline ER Approach

The goal of this paper is to investigate in how far multi-source ER tasks can be addressed using standard techniques. To this end, in this section we propose a baseline method that combines standard techniques in a standard ER process.

The typical ER process consists of four main steps (Christen, 2012; Naumann and Herschel, 2010) that are illustrated in Figure 1. We will briefly explain these phases and discuss the choices made in the baseline method.

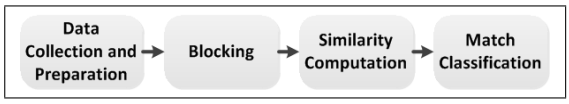


Figure 1: General Entity Resolution Process

3.1 Data Collection and Preparation

The first step is data collection and preparation during which the raw data is collected from various sources, then cleaned and preprocessed. During this step we have to assure that collected data has the same format (standardized date, null values, special characters, etc).

Therefore, we extract all person references from civil certificates. As shown in Table 1, this part of the data has a pre-formatted structure. Therefore, data extraction is a straightforward task. Table 3 shows three sample references which are extracted from the civil certificate of Table 1.

Table 3: The references extracted from the sample civil certificate in Table 1.

ref_ID	Person Name	Place	Date	Cert_ID
124358	Theodor Werners	Erp	14-04-1861	6453
124359	Peter Werners	-	-	6453
124360	Anna Meij	-	-	6453

The free-text available in notary acts requires an additional preprocessing step to extract the person references from them. Natural Language Processing (NLP) and Named Entity Recognition and Disambiguation tools (Chowdhury, 2003; Nadeau and Sekine, 2007) can be used for extracting informative features such as name entities and locations out of the text. In our case we applied the NLP tool Frog[1] (Van den Bosch et al., 2007) which is a Dutch morph-syntactic analyzer and dependency parser. A sample person reference extracted form the notary act of Table 2, using NLP software, is shown in Table 4.

[1] http://ilk.uvt.nl/frog/

Table 4: The reference extracted by means of the NLP tool and available data from the certificate (i.e., place and date).

ref_ID	Person Name	Place	Date	TextID
94254	Theodor Werners	Boekel	24-07-1896	100

As can be seen in Table 4 the data extracted from a text has only a few features as compared to the structured data shown in Table 1.

3.2 Blocking

The second step of the ER process is blocking. In order to avoid having to compare all pairs of references, for each reference we use a blocking key to split all references into different blocking partitions. It allows to reduce data complexity and diminishes the number of potential candidate reference pairs. A standard example of blocking keys are a phonetic encoding of name or last name, time period range or geographical distance. Although blocking techniques are commonly used in ER, there are other available methods such as using bit vectors (Schraagen, 2011) or hashing (Kim and Lee, 2010). They are not considered as blocking but also to reduce the number of potential candidate pairs.

Therefore, in order to avoid having to compare all pairs of references, we assign multiple blocking keys to each reference based on the phonetic encoding of first and last names. We follow the assumption that a name is often spelled in many different ways. We use multiple blocking keys such as *Double Metaphone encoding* (Philips, 2000) that take into account the various spelling and phonetic rules and *Soundex encoding* (Bourne and Ford, 1961) which indexes a name by sounds rather than spelling. Below is an example of applied blocking keys to encode imprecise names from Table 1 and Table 2.

Table 5: An example of blocking keys used in the baseline approach.

Name	Soundex	Double Metaphone
Teod**oo**r	T600	TTR
Theod**o**r	T600	TTR

3.3 Similarity Computation

During the similarity computation step the similarity score between two attributes, associated with two distinct references, is computed based on their types. For the attribute with type $String$ common similarity measures are character-based, token-based or phonetic measures, for instance *Levenshtein Edit distance*, *Jaro Winkler distance*, *Monge Elkan distance*, *Jaccard Coefficient*, *Cosine similarity*, *Soundex*, *Double Metaphone* and etc. (Elmagarmid et al., 2007; Winkler, 1995). For attributes with type $Date$ similarity can be calculated as date difference or as a binary value {*true, false*} that represents if two dates are the same or not.

For every pair of references in the same block we compare essential attributes using a suitable similarity measure. To be more precise, to compare the person names we use the similarity measure called *Jaro-Winkler* which return a number between 0 and 1, where 1 is the highest value when two names are exactly the same (Naumann and Herschel, 2010). Thus, for instance, the comparison of the two names $Teodoor$ and $Theodor$ with the Jaro-Winkler measure yields $0,91$.

We consider the similarity between dates and places as a boolean value which is $true$ when the date range between two references is plausible or places of two references are the same. We determine the plausibility of the date range based on internal rules, for example, the date in a birth certificate can't be earlier than 80 years before the date in a notary act, and the date in a death certificate can't be later than 80 years after the date in a notary act. We assume that maximum lifespan is 100 years.

In the baseline method we compute the similarity only between three attributes such as person name, date and place. Since those attributes are successfully extracted from a notary act.

3.4 Classification

The last step of the overall ER process is a classification. The score function computes the final similarity score between two references based on results of single comparison measures. There is a variety of techniques for designing a score function that combines individual similarity scores from statistics, modeling, machine learning and data mining (Florian et al., 2003; Christen, 2008). Many of them require a prior training phase on a representative subset of data to make a more efficient prediction on new data.

After that, pairs of references are classified into classes *Matched* or *non-Matched* based on a threshold value of the score function.

We learn the score function using a supervised learning technique and consider the genealogical ER problem as a problem of prediction. For learning the score function we use a training dataset that we will discuss in detail in the Section 4.3

We apply a linear scoring model as a predictive model and calculate the score function as follows:

$$\text{Score}(r_i, r_j) = w_0 + \sum_{l=1}^{k} w_l \cdot \text{sim}(r_i.a_l, r_j.a_l) \quad (1)$$

The parameters w_0 to w_k are learned in a training phase. The function $\text{sim}(r_i.a_l, r_k.a_l)$ represents similarity measures of the attribute a_l between two arbitrary references r_i and r_j, while reference r_i and r_j have k attributes in common.

4 Experiments

The application of the baseline ER approach and its evaluation on real-world data requires additional steps. The overall experimental setup is depicted in Figure 2. The first step is the process of gathering expert opinions. This is a crucial requirement for the evaluation of the baseline. Therefore, in Subsection 4.1 we present an interactive web-based interface which is used for getting input from human experts. Subsequently, in Subsection 4.2 we explain the classifier learning and the training set construction. Finally, we elaborate on the application and the evaluation of the model.

4.1 Manual labeling phase

In order to generate the adequate training/test set for classification process, a web-based interactive tool is developed (Efremova et al., 2013) which allows historians to navigate through the structured and unstructured data, and label the matches they find between various references. This tool is built using the Django[2] framework and uses various programming tools for storage, exploration and refinement of available data. It benefits from an intelligent searching engine, developed based on the Solr[3] enterprise search platform, with which historians can easily search through the dataset. Ba-

[2] https://www.djangoproject.com/

[3] http://lucene.apache.org/solr/

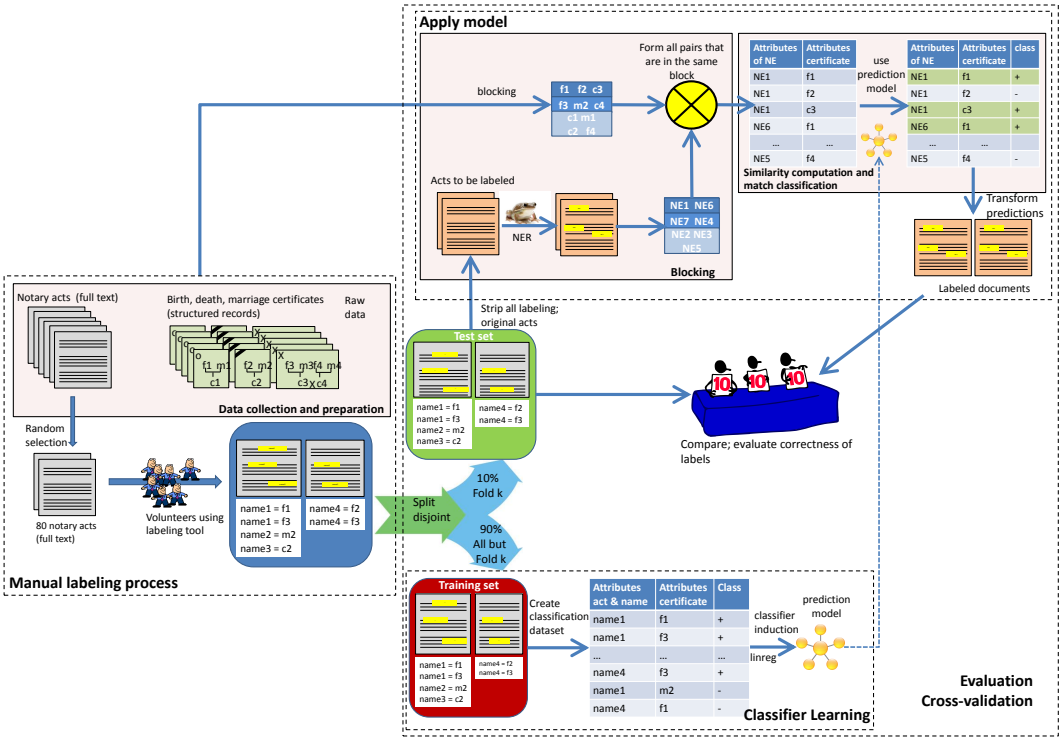


Figure 2: The training and validation process: A random subset of data is selected and human experts have labeled them manually. Afterward, the human labels are split into different partitions to construct training and test datasets.

sically, the required data can be found via person name, location, date and relations.

The developed Labeling tool, shown in Figure 3, is very powerful and easy to use, which assists historians to link name-references mentioned in notary acts to name-references mentioned in civil certificates.

The time required to report a correct match between two name-references varies from a few seconds to probably hours of time, depending on how similar two references are (e.g., whether places, dates, ages, professions and relatives match or not), and how easy it is to compare those two references. Consequently, the level of confidence in reporting a match varies. Therefore, the actions that historians take (e.g., which keywords they take and how fast they can recognize a match), and their level of confidence in reporting the match are all stored in the database. As a result, a rich benchmark is generated that includes the list of matches, the level of confidence and the list of actions that historians search for before reporting the match.

In this work, we consider each pair of references labeled by a historian as an example of a positive match between two different sources of data. Due to insufficient information in a notary act, incomplete civil certificates or a very frequent person name, no matches might be found for some references. We assign a zero-matched status to such references for which a corresponding match could not be found.

4.2 Classifier Learning

The output of the labeling procedure contains two types of labeled data: pairs of positive matches and zero-matched references. However, for the training process, any standard classifier requires a collection of positive and negative examples. For obtaining the negative examples we use random combinations of zero-matched references and other unrelated references.

Figure 2 depicts the process of learning a classifier on the training set.

4.3 Cross validation

In order to assess the performance of the ER process, we need an evaluation technique that excludes various biases and hypotheses. As gathering an extra validation dataset is very costly, we apply 10-fold cross-validation method on the entire baseline approach. We randomly partition the manually labeled data into 10 equal size subsets. Then one subset is chosen as the validation data for testing the classifier, and the remaining subsets

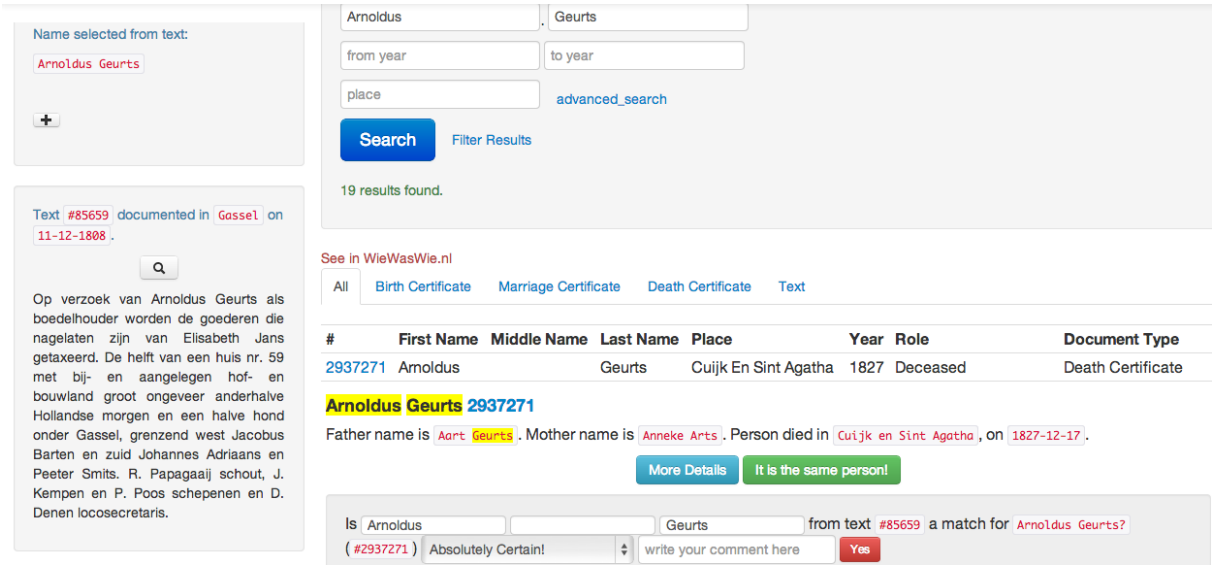


Figure 3: The developed web-based labeling tool for generating the required training/test dataset.

are used for training the classifier. Then the cross-validation process is repeated 10 times, with each of the 10 subsets used exactly once as the validation dataset. The 10 results from the folds then are averaged to produce a single estimation. To clarify that the overall process was cross-validated we include Figure 2.

5 Evaluation Results

In order to evaluate the performance of the baseline ER approach, we compute the sets of True Positives (TP), False Positives (FP) and False Negatives (FN) as the correctly identified, incorrectly identified and incorrectly rejected matches, respectively.

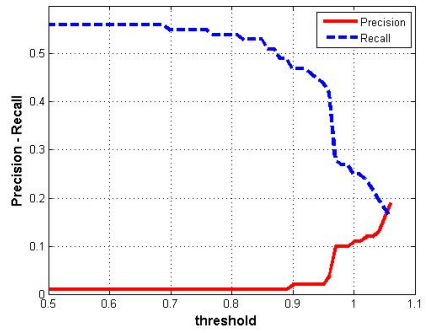
The precision and recall for different thresholds are shown in Figure 4a and Figure 4b. As shown in this figure, high precision is just achieved for very high thresholds; however, the recall is very low for these thresholds. A fast drop of precision with the decrease in threshold, and consequently increase in recall, is a drawback of the proposed baseline, which shows how important a deeper study of this problem and possible improvements are.

To show the insufficiency of the baseline approach, the F-measure is illustrated in Figure 5 for different similarity score thresholds. Although, increasing the threshold improves the matching performance, this improvement is not sufficient to make the automatic matching suitable for real-life applications.

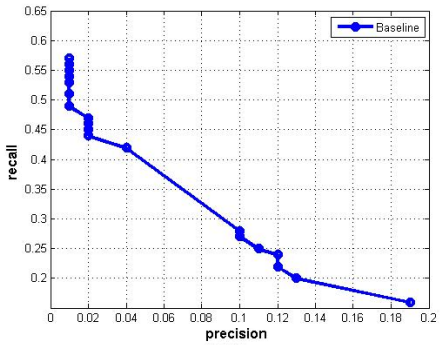
Alternative analysis. In this section in Table 5 we show a detailed comparative analysis of the number of matches identified by humans and by the baseline method with two selected threshold levels of score function $T = 1.05$ and $T = 1$. We follow the assumption that each person name extracted from the notary act should correspond to only one birth and one death certificate, the number of marriage certificates ranges from 1 to 3. Historians manually found 643 positive matches that correspond to 169 person entities and distributed between birth, marriage, death certificates and children documents where a person is mentioned as a parent (father or mother). The number of total manual matches is not very high because for every person extracted from a notary act, historians have to find an appropriate match among other 7,500,000 person references which is very difficult. As is obvious from Table 5, the historical information is not complete. The number of matches identified by historians between notary acts and civil certificates is very different from the expected number as well as the baseline outcomes.

Table 6: Number of matches according to humans and baseline approach. B, D, M stand for birth, death and marriage certificates. BC, MC, DC means matches identified in the certificates of children: birth, marriage and death respectively.

Method	B	D	M	BC	MC	DC
Expected	169	169	x	x	x	x
Humans	18	83	43	132	135	232
T = 1.05	5	39	6	21	28	62
T = 1	8	70	16	66	273	108



(a)



(b)

Figure 4: Evaluation of trained classifier. (a) Precision and Recall in terms of different matching score thresholds (b) Recall vs. Precision

6 Discussion

As can be seen from Section 5 the baseline approach requires significant improvements and the direct application of the standard ER solution to real-world multi-source genealogical dataset does not bring satisfactory results. Dealing with these issues will be topic our future research. We hope to improve the results by using the *relational information* of references. For example, if we find a couple (husband and wife) in a notary act that is also mentioned in a marriage certificate, then the probability that they are the same people increases. Another potential way to improve results is an extraction of extra features such as family relationships from notary acts with NLP/NER tools.

In addition, the lack of ground truth makes it difficult to get reliable and high-quality evaluation. The data obtained during the manual labeling process, was considered as a ground truth. Nevertheless, the real ground truth remains unknown and may differ from the human judgment. The difficulty is that based on the available information, humans often cannot conclude whether persons mentioned in the text documents and in civil

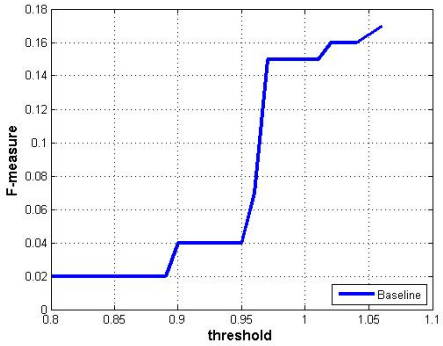


Figure 5: F-measure value in terms of increase in similarity score threshold values

certificates are the same or not. Since we do not know an absolute ground truth, we use alternative way and consider available manual labeling as a ground truth. In Figure 6 using Venn diagram, we demonstrate all possible intersections when a match is positive according to the absolute ground truth (GT), the human judgment, and the baseline approach.

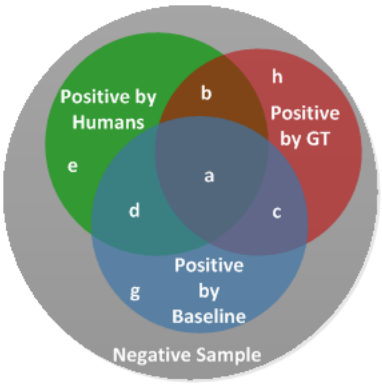


Figure 6: A diagram of possible intersections between the ground truth, human judgment and the baseline approach

Each circle in the diagram represents positive matches according to absolute ground truth, human judgment and the baseline approach. The closer human judgment agrees with the absolute ground truth, the more accurate is our evaluation. Because of having e, d, c and h subsets we always overestimate or underestimate the true positive rate (TPR) and the false positive rate (FPR). In our case it is sometimes impossible to obtain the ground truth, but the combination of human efforts, automatic methods, and alternative analysis of results will help to make correspond the human judgment with the ground truth.

7 Conclusion

In this paper we studied the concept of ER in genealogical data research, where the data is provided from sources with different structures. Considering the multi-source characteristic of the data, the classical ER techniques are not directly applicable due to the diverse types of data attributes. Therefore, in this study a baseline approach, inspired by classical ER techniques was proposed, which uses the available common attributes in all data sources. In order to assess the effectiveness of the baseline ER approach, an interactive web-based labeling tool was developed with which the human experts helped to manually identify the matches from an adequate sample of the whole data. The manually labeled matching was used both for training the ER baseline approach, and computation of precision, recall, and F-score.

As future work, the authors are working on more advanced ER techniques based on Probabilistic Relational techniques, which can incorporate other available features in data such as relations between references, typing errors, etc.

8 Acknowledgments

The authors are grateful to the BHIC Center for the support in data gathering, data analysis and direction. In particular, we would like to thank Rien Wols and Anton Schuttelaars whose efforts were instrumental to this research and their patience and support appeared infinite.

References

Gerrit Bloothooft. 1995. Multi-source family reconstruction. *History and Computing*, pages 90–103.

Charles P. Bourne and Donald F. Ford. 1961. A study of methods for systematically abbreviating English words and names. *J. ACM*, 8(4):538–552.

Gobinda G. Chowdhury. 2003. Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89.

Peter Christen. 2008. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD International conference on Knowledge Discovery and Data mining*, KDD '08, pages 151–159, USA. ACM.

Peter Christen. 2012. *Data matching*. Springer Publishing Company, Incorporated.

Julia Efremova, Bijan Ranjbar-Sahraei, Frans A. Oliehoek, Toon Calders, and Karl Tuyls. 2013. An interactive,web-based tool for genealogical entity resolution. In *25th Benelux Conference on Artificial Intelligence*, pages 376–377, The Netherlands.

Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 168–171, USA. Association for Computational Linguistics.

Lise Getoor and Ashwin Machanavajjhala. 2012. Entity resolution: Theory, practice & open challenges. In *International Conference on Very Large Data Bases*, pages 2018–2019.

Steve Ivie, Graham Henry, Haven Gatrell, and Christophe Giraud-Carrier. 2007. A metric-based machine learning approach to genealogical record linkage. In *In Proceedings of the 7th Annual Workshop on Technology for Family History and Genealogical Research*.

Hung Kim and Dongwon Lee. 2010. Harra: fast iterative hashed record linkage for large-scale data collections. In Ioana Manolescu, Stefano Spaccapietra, Jens Teubner, Masaru Kitsuregawa, Alain Lger, Felix Naumann, Anastasia Ailamaki, and Fatma zcan, editors, *EDBT*, volume 426 of *ACM International Conference Proceeding Series*, pages 525–536. ACM.

John S. Lawson. 2006. Record linkage techniques for improving online genealogical research using census index records. In *Proceeding of the Section on Survey Research Methods*, pages 3297–3302.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher: John Benjamins Publishing Company.

Felix Naumann and Melanie Herschel. 2010. *An introduction to duplicate detection*. Morgan and Claypool Publishers.

Lawrence Philips. 2000. The double metaphone search algorithm. *C/C++ Users J.*, 18(6):38–43.

Marijn Schraagen and Hendrik J. Hoogeboom. 2011. Predicting record linkage potential in a family reconstruction graph. In *23th Benelux Conference on Artificial Intelligence(BNAIC'11)*, pages 199–206, Belgium.

Marijn Schraagen. 2011. Complete coverage for approximate string matching in record linkage using bit vectors. In *ICTAI'11*, pages 740–747.

Cary Sweet, Tansel Özyer, and Reda Alhajj. 2007. Enhanced graph based genealogical record linkage. In *Proceedings of the 3rd international conference on Advanced Data Mining and Applications*, ADMA07, pages 476–487. Springer-Verlag.

Antal Van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for dutch. In *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, pages 99–114.

William E. Winkler. 1995. Matching and record linkage. In *Business Survey Methods*, pages 355–384. Wiley.