Severe Name Homonymy: A Learning to Rank Approach for Civil Registry Entity Linking

Marco T. Dutra, Lucas G. L. Costa, Felipe D. S. Martins, Gabriela M. M. Freitas, Leonardo C. C. Castro, Pedro G. B. Vieira, Michele A. Brandão, Gisele L. Pappa

[1] Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

{marco.dutra, lucas.lage, felipedias, gabrielamiserani, leonardocesar pedrobarruetavena, michele.brandao, glpappa}@dcc.ufmg.br

Abstract. *This study addresses the challenge of transforming string-based kinship references into explicit entity relationships within large-scale civil registries. In scenarios where a record provides only the name of a relative (e.g., a mother's name), identifying the correct unique identifier among numerous homonymous candidates constitutes a complex Known-Item Search problem. We propose a supervised Learning to Rank framework designed to disambiguate and rank these candidates by integrating kinship-based relational features, conditional probability signals, demographic indicators, and BERT-based semantic similarity. The approach was evaluated on a large-scale private dataset from Brazil and four public datasets: IPUMS (covering USA, Norway, and Iceland) and Wikidata. Experimental results demonstrate that our method provides a robust solution for identity resolution under high name ambiguity, achieving a Recall@5 above 0.90 across all evaluated datasets and enabling the construction of accurate entity links in civil databases.*

1. Introduction

Civil registries store structured information about individuals, typically including a unique identifier, a full name and demographic information [Nations 2014]. In many cases, these records also contain references to relatives, such as parents, spouses, or children [Ruggles et al. 2025b]. However, while the primary individual in a record is associated with a unique identifier, the referenced relatives frequently appear only by name, without an explicit identifier, which makes linking records a challenging task.

In governmental auditing and investigation scenarios, establishing reliable links between individuals (i.e., resolving textual name references to concrete entities through their unique identifiers) is essential. This task can be formulated as an Entity Linking problem [Shen et al. 2014] over structured civil data. Given a target individual and the name of a relative extracted from a civil record, the goal is to identify the correct person in the registry and recover their unique identifier.

A straightforward name-based retrieval strategy is typically insufficient, as names are not unique identifiers and often lead to ambiguity in large populations [Song et al. 2025]. Distinguishing the correct relative from homonymous individuals therefore becomes a central difficulty of the task, that often requires sophisticated disambiguation strategies beyond simple name matching [Zhou et al. 2024].

This study aims to support investigators and public-sector auditors in reconstructing kinship networks of persons of interest. Rather than performing fully automatic clas-

sification of the correct relative among theirs homonymous, our goal is to produce a high-quality ranking of candidates, ideally placing the true relative at the top positions. In investigative and auditing workflows, final confirmation is performed manually by domain experts, as these processes demand a high degree of reliability. Consequently, our objective is to provide a ranking mechanism that meaningfully reduces manual effort while preserving human oversight in the final decision.

We formulate this problem as a Known-Item Search [Schütze et al. 2008] task for Entity Linking in the presence of name homonymy, where the target entity is known to exist in the database but must be retrieved and ranked among multiple candidates sharing the same name. To address this challenge, we propose a supervised Learning to Rank [Liu 2009] framework tailored to Known-Item Search in civil registries. The approach leverages kinship-based relational features, conditional probability signals, semantic similarity scores derived from a BERT-based language model [Zhang et al. 2019], and demographic indicators to prioritize the correct individual among competing candidates.

We validate the proposed methodology on a large real-world private civil registry dataset from Brazil and on four public datasets: IPUMS [Ruggles et al. 2025b, Ruggles et al. 2025a] (USA, Norway, and Iceland) and Wikidata [Vrandecic and Krötzsch 2014]. Across datasets with different languages, schemas, and population characteristics, the results show that Learning to Rank for Known-Item Search provides a scalable and robust solution for Entity Linking in civil registries under high levels of name ambiguity. In light of these results, this paper makes the following contributions:

1. Problem Formalization. We formally define kinship-based Entity Linking in civil registries as a Known-Item Search problem under name homonymy.

2. Ranking-Based Approach. We introduce a supervised Learning to Rank approach that integrates demographic consistency signals, relational constraints derived from family structure, and contextual attributes to prioritize the correct individual among homonymous candidates.

3. Cross-Dataset Empirical Validation. We conduct extensive empirical validation across heterogeneous datasets, including a large real-world Brazilian civil registry and multiple public datasets (IPUMS USA, Norway, Iceland, and Wikidata), demonstrating robustness and generalizability across languages, schemas, and population characteristics.

The remainder of this paper is structured as follows. Section 2 reviews the related work. Section 3 formally defines the problem setting. Section 4 describes the proposed approach. Section 5 presents the datasets, experimental design, and empirical results. Finally, Section 6 concludes the paper by discussing key findings, limitations, and directions for future research.

2. Related Work

Entity Linking is a fundamental task for integrating entities across heterogeneous data sources, such as Wikidata, as evidenced by recent systematic literature reviews [Scharpf et al. 2026]. When formulated as a Learning-to-Rank task, Entity Linking has increasingly leveraged transformer-based language models to perform the ranking step as a core component of the learning process [Wu et al. 2020, Mrini et al. 2022].

However, not all approaches adopt a ranking perspective. Methods such as E-BELA [Ítalo Pereira and Ferreira 2024] also employ language models, but primarily frame Entity Linking as a classification problem and evaluate performance using accuracy rather than ranking-based metrics. Furthermore, [Albuquerque et al. 2025] demonstrates that combining language models with domain knowledge and efficient indexing strategies can lead to more robust and scalable solutions.

Beyond Entity Linking, similar challenges arise in Record Linkage [Christen 2012], where the goal is to identify matching records across datasets. While our approach is framed as an Entity Linking task, it shares important similarities with Record Linkage, particularly in the need to resolve ambiguity among homonymous candidates using auxiliary attributes such as relational features and demographic indicators. Several studies have explored machine learning methods [Santos and Nascimento 2023, Price et al. 2021, Ali Omar et al. 2023] and Transformer-based models [Arora and Dell 2024, Ornstein 2025] for Record Linkage.

In the specific context of historical civil registries and population data, prior work has proposed matching methods tailored to these domains. For instance, [Nanayakkara and Christen 2022] utilized locality-sensitive hashing with spatial and temporal constraints to link population records across different periods. Similarly, [Mon Myint and Naing 2023] introduced a graph-based household matching method for historical census data, leveraging the structural context of family units to improve linkage.

A closely related problem that also involves name homonymy is author name disambiguation in academic databases. Recent approaches typically formulate this task as a global clustering problem, grouping records belonging to the same individual based on signals such as co-authorship networks and textual similarity [Zhou et al. 2024, Tian et al. 2024]. In contrast, our setting assumes a single registry and a known target individual. Instead of global clustering, we formulate the task as Entity Linking via Known-Item Search, where candidates are retrieved based on name homonymy and subsequently ranked to identify the correct match.

3. Problem Formalization

This section presents the problem formalization, Figure 1 provides an overview of the proposed formulation.

Let $P = \{p \mid p \text{ is a person in the civil registry}\}$ be the set of all individuals. Each person $p \in P$ is associated with a set of attributes, including a unique identifier p[id], full name p[name], date of birth p[birth_date], a set of addresses p[addrs], a set of occupations p[occs], and textual references to parents, namely p[father_name] and p[mother_name].

Relatives without Identifiers. For a given person $p \in P$, let r_p denote a relative of p whose identifier is unknown. Formally, r_p is defined only by its name: $r_p[name] \neq \varnothing$ and $r_p[id] = \varnothing$. Let E be the set of possible kinship relation types, $E = \{e \mid e \text{ is a kinship relation type}\}$. For each pair (p, r_p), there exists a relation type $e_{p,r} \in E$ indicating the family relationship between p and r_p.

Homonymous Candidate Set. Given $r_p[name]$, we define the homonymous candidate set: $H_r = \{h \in P \mid h[name] = r_p[name]\}$. That is, H_r contains all individuals in the registry whose name matches the relative name. We assume a Known-Item Search

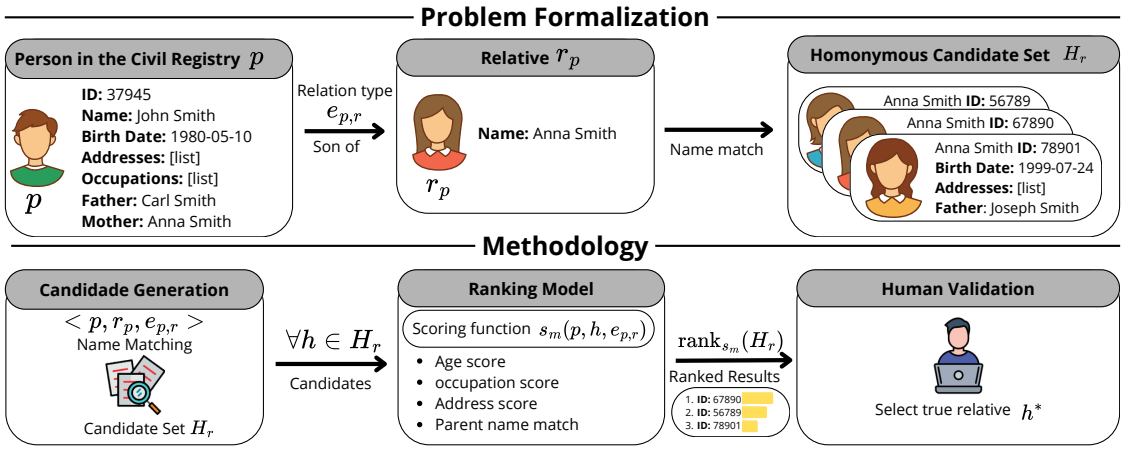

Figure 1. Overview of the problem formalization and the overall approach.

setting, meaning that there exists exactly one true entity $h^* \in H_r$, such that $h^*[id] \neq \varnothing$ and h^* corresponds to the real-world relative of p.

Objective. The task is to recover the unknown identifier $r_p[id]$ by identifying h^* within H_r. Let $s : H_r \to \mathbb{R}$ be a scoring function that assigns a relevance score to each candidate. Define the ranking $\text{rank}_s(H_r)$ as the ordering of H_r induced by sorting candidates in decreasing order of s. The goal of the proposed method m is to learn a scoring function s_m such that $\text{rank}_{s_m}(h^*)$ is minimized, i.e., the true relative h^* appears as high as possible in the ranked list.

Entity Linking Decision. The ranked list $\text{rank}_{s_m}(H_r)$ is presented to a human analyst, who performs the final confirmation step by selecting the correct entity h^*.

4. Methodology

This section presents the proposed methodology. Figure 1 provides an overview of the problem formalization and the overall approach. Section 4.1 describes the learning setting, while Section 4.2 details the feature engineering process. Finally, Section 4.3 presents the inference pipeline.

4.1. Learning Setting

The objective of the proposed method m is to learn a scoring function s_m such that $\text{rank}_{s_m}(h^*)$ is minimized, i.e., the true relative h^* appears as high as possible in the ranked candidate list.

We adopt a supervised Learning to Rank approach under a Known-Item Search setting. Let $D = \{\langle p, r_p, e_{p,r}, H_r \rangle\}$ be the dataset, where each instance consists of: $p \in P$ is the reference person; r_p is the partially observed relative; $e_{p,r} \in E$ is the kinship relation type and H_r is the homonymous candidate set.

The dataset D is split into D_{train}, $D_{\text{validation}}$, D_{test}. To enable supervised training, each candidate set H_r is expanded into pairwise instances $\langle p, h, e_{p,r}, y_{p,h} \rangle$, $\forall h \in H_r$, where $y_{p,h} = 1$ if $h = h^*$ (i.e., the true match) and $y_{p,h} = 0$ otherwise. Thus, the ranking problem is reduced to learning a scoring function $s_m(p, h, e_{p,r})$ that assigns higher scores to true relatives.

Given the feature vector $\phi(p, h, e_{p,r})$, the model m learns a scoring function $s_m(p, h, e_{p,r}) = f_\theta(\phi(p, h, e_{p,r}))$ using a supervised learning-to-rank algorithm trained on pairwise labeled instances. The training objective follows standard learning-to-rank formulations that optimize ranking quality metrics by encouraging correct candidates to be ranked higher than incorrect ones. Formally, it enforces $s_m(p, h^*, e_{p,r}) > s_m(p, h, e_{p,r})$ for all $h \neq h^*$. At inference time, the model produces a ranked candidate set $\text{rank}_{s_m}(H_r)$, where candidates h are ordered according to their predicted scores $s_m(p, h, e_{p,r})$.

4.2. Feature Engineering

For each pair (p, h), we construct a feature vector $\phi(p, h, e_{p,r})$ composed of demographic, relational, and semantic compatibility signals, as detailed below.

Address. Let $p[\text{addrs}]$ and $h[\text{addrs}]$ denote the sets of addresses associated with individuals p and h. We compute the maximum pairwise semantic similarity using a BERT-based measure:

$$\text{AddrScore}(p, h) = \max_{a_p \in p[\text{addrs}],\ a_h \in h[\text{addrs}]} \text{BERTScore}(a_p, a_h) \tag{1}$$

When structured address data is available, we complement this score with exact matching features over country, state, city, neighborhood, and street number, denoted by $\text{AddrExt}(p, h)$, and Levenshtein-based similarity over street names and address complements, denoted by $\text{AddrLev}(p, h)$.

Age. Let $\Delta_{\text{age}}(p, h) = |p[\text{birth\_date}] - h[\text{birth\_date}]|$ denote the absolute age difference. For each relation type $e \in E$, we estimate the parameters of a Students t-distribution from the training set D_{train}, such that $\nu_e, \mu_e, \sigma_e = \text{fit}_t\big(\{\Delta_{\text{age}}(p, h) \mid (p, h) \in D_e\}\big)$, where $D_e \subseteq D_{\text{train}}$ contains all pairs with relation type e. Here, ν_e denotes the degrees of freedom, μ_e the location (mean), and σ_e the scale (standard deviation). The score is then computed as a normalized density:

$$\text{AgeScore}(p, h, e) = \frac{t_{\nu_e}\big(\Delta_{\text{age}}(p, h); \mu_e, \sigma_e\big)}{t_{\nu_e}\big(\mu_e; \mu_e, \sigma_e\big)} \tag{2}$$

Where $t_\nu(x; \mu, \sigma)$ denotes the probability density function of the Students t-distribution. This normalization ensures that the score is bounded in $[0, 1]$, with value 1 at the mode of the distribution, and captures how plausible the observed age difference is given the relation type.

Occupation. Let $p[\text{occs}]$ and $h[\text{occs}]$ denote the sets of occupations. We compute the maximum pairwise semantic similarity using a BERT-based measure:

$$\text{OccScore}(p, h) = \max_{o_p \in p[\text{occs}],\ o_h \in h[\text{occs}]} \text{BERTScore}(o_p, o_h) \tag{3}$$

Relation Type. The kinship type $e_{p,r}$ is included as a categorical feature ($\text{Enc}(e_{p,r})$).

Parental Name Consistency. We define two relational indicators based on Levenshtein similarity: (i) $\text{ParPar}(p, h)$: similarity between the parental names of p and h; (ii) $\text{NamePar}(p, h)$: similarity between $p[name]$ and the father's or mother's name of h.

Formally, we define:

$$\phi(p, h, e_{p,r}) = [\ \text{AddrScore}(p, h), \text{AddrExt}(p, h), \text{AddrLev}(p, h), \text{AgeScore}(p, h, e_{p,r}), \\ \text{OccScore}(p, h), \text{Enc}(e_{p,r}), \text{ParPar}(p, h), \text{NamePar}(p, h)] \tag{4}$$

4.3. Inference Pipeline

At inference time, the process comprises three steps: (i) *Candidate Generation*, where the candidate set H_r is retrieved via name matching; (ii) *Ranking* (Known-Item Search), in which $s_m(p, h, e_{p,r})$ is computed for all $h \in H_r$ to produce the ranked set $\text{rank}_{s_m}(H_r)$; and (iii) *Entity Linking with Human Validation*, where the ranked list is presented to a domain expert for final confirmation of the correct entity h^*.

5. Experimental Evaluation

This section presents the experimental evaluation. Section 5.1 describes the datasets and preprocessing steps, while Section 5.2 details the instantiation of the proposed method. Section 5.3 introduces the baseline methods, and Section 5.4 outlines the evaluation setup. Finally, Section 5.5 reports and discusses the results. The complete experimental code for the public datasets is available in our GitHub repository[1].

5.1. Datasets

We conduct experiments on five datasets ($|D| = 5$), including one large-scale private dataset from the complete Brazilian Civil Registry, three public census datasets from the *IPUMS* project [Ruggles et al. 2025b, Ruggles et al. 2025a], and a dataset derived from Wikidata [Vrandecic and Krötzsch 2014]. Table 1 summarizes their main statistics and candidate set distributions. The set of kinship relations is denoted by $E = \{$*parent*, *child*, *sibling*, *spouse*, *grandparent*, *grandchild*, *aunt/uncle*, *nephew/niece*, *cousin*, *in-law*, *step/adopted*, *other*$\}$. The complete list of relations and their frequency per dataset are available in our GitHub repository[1].

The *IPUMS* (Integrated Public Use Microdata Series) provides harmonized census microdata across countries and historical periods, including individual and household level records with demographic attributes and explicit family relationships defined with respect to the household head. We restrict our analysis to historical samples (pre-1950), as these are the only datasets that include personal names, which are essential for homonym ranking. Additionally, we focus on Iceland, Norway, and the United States, as these datasets provide the complementary attributes required for our feature construction, namely age, address, and occupation.

D_1: Brazilian Civil Registry. A large-scale, real-world dataset containing demographic, relational, and contextual information. This dataset represents the most challenging scenario due to its scale and population diversity, resulting in a high prevalence of homonyms and a heavy-tailed candidate distribution. It includes complete and structured address information, making it a realistic benchmark for entity linking tasks. Notably, it is the only dataset that provides *Parental Name Consistency* features, enabling the exploitation of familial naming patterns for disambiguation.

[1] `https://github.com/MarcoTuMD/name_homonymy_ltr`

Table 1. Datasets statistics. Base Pop. and Exp. Pop. denote the number of individuals in the full dataset and experimental subset, respectively; Pairs is the number of kinship pairs (p, r_p). $\mu(|H_r|)$, $\sigma(|H_r|)$, $P_{50}(|H_r|)$, and $P_{90}(|H_r|)$ denote the mean, standard deviation, median, and 90th percentile of $|H_r|$.

Dataset	Base Pop.	Exp. Pop.	Pairs	$\mu(\lvert H_r\rvert)$	$\sigma(\lvert H_r\rvert)$	$\mathrm{P}_{50}(\lvert H_r\rvert)$	$\mathrm{P}_{90}(\lvert H_r\rvert)$
BR Civ. Reg.	280,509,059	17,938,638	279,406	460.84	3,039.26	10.06	487.00
Iceland	8,072	5,887	2,513	16.45	33.62	4.00	34.00
Norway	2,294,599	1,395,798	783,882	178.00	534.09	21.00	447.00
USA	6,103,822	2,011,666	1,155,657	14.21	41.21	3.00	32.00
Wikidata	12,963,076	531,296	264,507	10.69	47.39	2.00	98.30

$\mathbf{D}_2$: IPUMS Iceland (1729). [Garðarsdóttir nd]. Sample 352172901. A small-scale historical dataset with limited population and relationships, leading to low ambiguity. Address information is only available at the city level.

$\mathbf{D}_3$: IPUMS Norway (1900). [Archive et al. 2008]. Sample 578190001. A dataset with a higher frequency of homonyms due to naming conventions, resulting in a long-tailed candidate distribution. For example, patronymic surnames such as *Olsen* ("son of Ole") and *Johansen* ("son of Johan") are extremely common, leading to many individuals sharing identical full names. However, address information is limited to the city level.

$\mathbf{D}_4$: IPUMS United States (1930). [Ruggles et al. 2025b]. Sample 193002. A 5% population sample, since the complete census records are not available through IPUMS. Address information is available at the street level, making it highly discriminative, as few homonyms share the same street.

$\mathbf{D}_5$: Wikidata (English Subset). [Vrandecic and Krötzsch 2014] A collaborative knowledge base containing interlinked entities and explicit relations. Unlike civil registries, it focuses on notable individuals, resulting in fewer homonyms and more distinctive attributes. Address information is unstructured (free-text), e.g., "Malibu Beach" or "Manhattan Island." Data collection was performed through the WDumper tool[2], based on the official dump obtained on January 19, 2026.

Preprocessing. For each dataset D_i, we constructed a fully labeled ranking dataset according to the formalization described in Section 4. To ensure the validity of the instances, we enforced two constraints. First, each relative must have at least one homonymous candidate, i.e., $|H_r| > 0$. Second, at least one of those pairwise feature (AddrScore, AgeScore, OccScore) must be computable between the reference person p and a candidate h. After applying these constraints, each relative–candidate list was transformed into pairwise training instances of the form $\langle p, h, e_{p,r}, y_{p,h} \rangle$.

5.2. Instantiation of the Proposed Method

Recall from Section 4 that the objective is to learn a scoring function $s_m : H_r \to \mathbb{R}$, such that $\mathrm{rank}_{s_m}(h^*)$ is minimized. In this experimental evaluation, we instantiate the proposed method m using two gradient-boosted learning-to-rank algorithms:

$\mathbf{m}_1$: LGBMRanker. [Ke et al. 2017] We use the `lambdarank` objective, a pairwise ranking formulation that approximates the optimization of ranking metrics through

[2]WDumper: `https://github.com/bennofs/wdumper`

Table 2. Hyperparameter configuration of the ranking models.

Model	Objective	Num estimators	LR	Max Depth	Num Leaves	Sub sample	Col. sample	Tree Method	Seed
XGB Ranker	rank ndcg	300	0.05	6	–	0.8	0.8	hist	42
LGBM Ranker	lambda rank	300	0.05	-1	31	0.8	0.8	–	42

gradient-based updates.

m_2: XGBoostRanker. [Chen and Guestrin 2016] We use the `rank:ndcg` objective, which directly optimizes ranking quality with respect to NDCG.

Both models learn a scoring function $s_{m_i}(p, h, e_{p,r})$ from the pairwise feature representation described in Section 4. The predicted score is then used to rank candidates within each homonymous set H_r. Table 2 summarizes the hyperparameter settings used for each ranking model.

Class Imbalance Handling. To mitigate class imbalance during training, we applied negative sampling with a maximum ratio of 1:10. For each positive pair ($y_{p,h} = 1$), at most 10 negative candidates ($y_{p,h} = 0$) were included in the training set.

BERT Model. In Section 4.2, BERT-based similarity scores [Zhang et al. 2019] are used to compute *OccScore* and *AddrScore*. In our experiments, we employ the base version of *mmBERT* [Marone et al. 2025], a multilingual variant of ModernBERT. This model is chosen for its strong performance and its coverage of all languages present in our datasets, namely Portuguese, English, Norwegian, and Icelandic.

5.3. Baselines

m_3: Logistic Regression. We implemented a supervised binary classification baseline using logistic regression. The model is trained on pairwise instances $\langle p, h \rangle$ and produces probabilistic scores used for ranking. It uses L2 regularization (LBFGS solver, max_iter=3000) and class weighting inversely proportional to class frequencies. This baseline serves to demonstrate that a standard classification approach is insufficient to fully capture the ranking structure of the problem. It is trained and evaluated under the same cross-validation protocol as the ranking models.

m_4: Heuristic. We also implemented a deterministic rule-based strategy defined by experienced investigators specialized in relationship analysis and entity resolution. The heuristic uses training data solely to estimate the age distribution statistics required for AgeScore. It is evaluated on the same test splits as the learning-to-rank models. The heuristic combines similarity scores derived from age, address, and occupation features:

$$m_4 = 0.5 \cdot \text{AgeScore}(p, h, e_{p,r}) + 0.4 \cdot \text{AddrScore}(p, h) + 0.1 \cdot \text{OccScore}(p, h) \quad (5)$$

5.4. Evaluation Setup

Cross-Validation Protocol. Experiments were conducted using 5-fold cross-validation, yielding five runs. Folds were created at the reference person level (p). In each run, one

Table 3. Results across all models and datasets.

Dataset	Metric	Heuristic	Log. Reg.	XGBoost	LGBM
BR Civ. Reg.	MRR	0.663 ± 0.006	0.741 ± 0.059	0.849 ± 0.013	**0.853 ± 0.012**
	Recall@1	0.572 ± 0.008	0.661 ± 0.073	0.799 ± 0.017	**0.804 ± 0.016**
	Recall@5	0.771 ± 0.005	0.836 ± 0.042	0.908 ± 0.008	**0.909 ± 0.008**
Iceland	MRR	0.754 ± 0.108	0.871 ± 0.012	0.906 ± 0.007	**0.906 ± 0.004**
	Recall@1	0.626 ± 0.140	0.782 ± 0.020	**0.838 ± 0.013**	0.833 ± 0.006
	Recall@5	0.916 ± 0.063	0.981 ± 0.007	0.986 ± 0.004	**0.986 ± 0.002**
Norway	MRR	0.463 ± 0.002	0.874 ± 0.000	0.878 ± 0.000	**0.879 ± 0.000**
	Recall@1	0.335 ± 0.002	0.798 ± 0.001	0.805 ± 0.000	**0.806 ± 0.001**
	Recall@5	0.609 ± 0.002	0.968 ± 0.001	0.970 ± 0.001	**0.971 ± 0.001**
USA	MRR	0.763 ± 0.000	0.992 ± 0.000	0.994 ± 0.003	**0.994 ± 0.000**
	Recall@1	0.645 ± 0.001	0.985 ± 0.000	0.986 ± 0.000	**0.987 ± 0.000**
	Recall@5	0.910 ± 0.000	0.999 ± 0.000	**0.999 ± 0.000**	**0.999 ± 0.000**
Wikidata	MRR	0.892 ± 0.001	0.898 ± 0.001	0.927 ± 0.000	**0.928 ± 0.001**
	Recall@1	0.822 ± 0.002	0.832 ± 0.002	0.880 ± 0.001	**0.881 ± 0.001**
	Recall@5	0.975 ± 0.001	0.978 ± 0.001	0.984 ± 0.000	**0.985 ± 0.000**

fold (20%) was used for testing, while the remaining four folds (80%) were split into training (90%) and validation (10%). Pairwise instances were generated after the split to prevent data leakage. Results are reported as the mean across the five runs.

Evaluation Metrics. We evaluate ranking quality using Recall@{1, 5} and Mean Reciprocal Rank (MRR). Since the task is formulated as a Known-Item Search problem, each kinship pair (p, r_p) is associated with a single relevant entity h^*. In this setting, precision-based metrics are inherently limited, as Precision@5, for instance, has a maximum value of 0.2. Consequently, MRR is particularly suitable, as it directly captures how early the true entity h^* appears in the ranked list.

5.5. Results

We evaluate four methods across five datasets using 5-fold cross-validation. Table 3 and Figure 2 report mean performance and standard deviation. Across all datasets and metrics, *Heuristic* consistently yields the lowest performance, followed by *Logistic Regression*, while both learning-to-rank models substantially outperform the baselines. This pattern indicates that neither heuristic nor classification-based approaches are sufficient for the task, highlighting the importance of explicitly optimizing a ranking objective. The baselines exhibit higher variance, particularly *Heuristic* on *IPUMS Iceland*, where results are sensitive to data splits due to the dataset's small size, and *Logistic Regression* on the *Brazilian Civil Registry*, as it cannot capture non-linear relationships. In contrast, *XGBRanker* and *LGBMRanker* show consistently low variance, indicating greater robustness and stability.

Although both learning-to-rank models achieve similar mean performance, we assess statistical significance using the Wilcoxon signed-rank test, appropriate for paired, non-normal samples. For each dataset and metric, we test whether the median difference between models is zero at $\alpha = 0.05$. And to further analyze model behavior, Table 4 reports SHAP-based relative feature importance percentages [Lundberg and Lee 2017, Lundberg et al. 2020] for *LGBMRanker*, averaged across folds. We focus on *LGBMRanker* as it achieves the best overall performance, as discussed below.

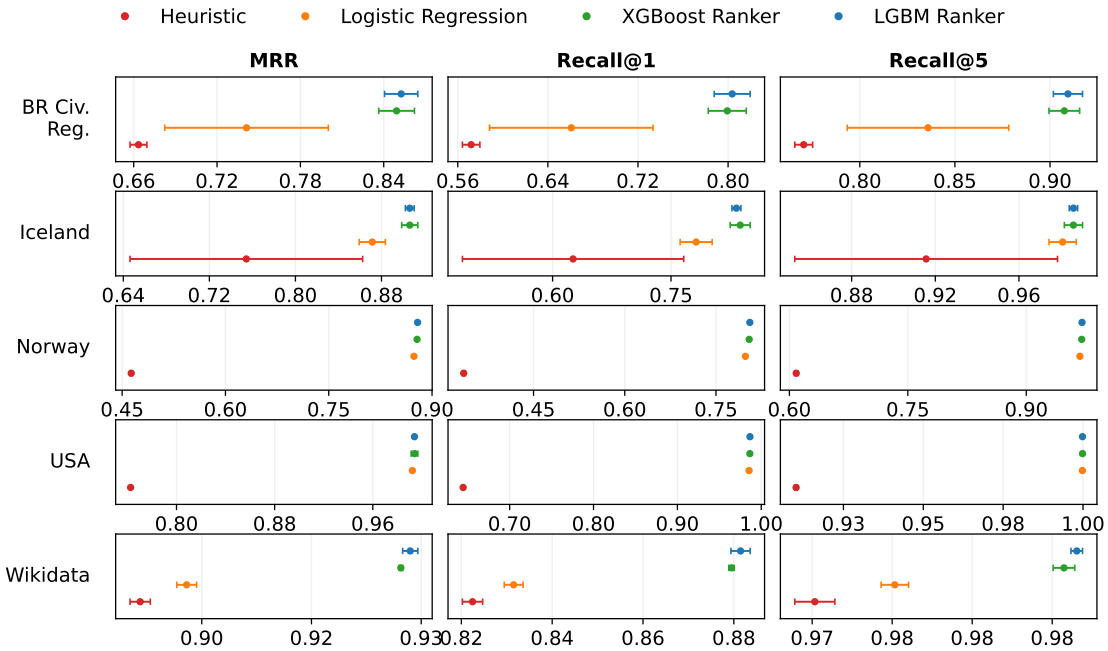

Figure 2. Graphical comparison of results across models and datasets.

Table 4. SHAP-based relative feature importance for *LGBMRanker*.

Dataset	Address (%)	Age (%)	Occup. (%)	Relation (%)	Par. Name (%)
BR Civ. Reg.	25.76 ± 0.27	19.68 ± 0.23	8.63 ± 0.31	27.28 ± 0.53	18.70 ± 0.36
Iceland	63.57 ± 2.07	20.78 ± 1.74	11.96 ± 0.74	3.69 ± 0.97	–
Norway	59.20 ± 0.92	18.90 ± 0.81	20.78 ± 1.17	1.12 ± 0.34	–
USA	84.52 ± 1.10	12.18 ± 1.26	1.67 ± 0.16	1.63 ± 0.20	–
Wikidata	14.81 ± 0.84	66.04 ± 0.89	14.42 ± 0.39	4.72 ± 0.26	–

Brazilian Civil Registry. *LGBMRanker* outperforms *XGBRanker* across all metrics ($W = 15$ and $p \leq 0.05$). The *LGBMRanker* model achieves strong performance, with mean values above 0.85 for MRR, 0.80 for Recall@1, and 0.90 for Recall@5. Its feature importance distribution is relatively balanced, indicating that multiple signals jointly contribute to the ranking process, which reflects the intrinsic complexity of the dataset where no single feature is sufficient for reliable disambiguation. This dataset constitutes the most challenging scenario due to its scale, population diversity, and real-world nature, resulting in a high prevalence of ambiguous cases such as homonyms and providing a realistic benchmark for deployment conditions.

IPUMS Iceland. *LGBMRanker* and *XGBRanker* achieve nearly identical performance across all metrics, with no statistically significant differences (MRR: $W = 12$, $p > 0.05$; Recall@1: $W = 11.5$, $p > 0.05$; Recall@5: $W = 11.5$, $p > 0.05$). This dataset is small, with relatively few individuals and relationships, which simplifies the task and reduces ambiguity. As a result, the *LGBMRanker* model achieve strong performance, with mean values above 0.90 for MRR, 0.83 for Recall@1, and 0.98 for Recall@5. Feature importance analysis indicates a dominance of address-related features (63.57%), while age remains a relevant signal (20.78%).

IPUMS Norway. *LGBMRanker* outperforms *XGBRanker* across all metrics ($W = 15$ and $p \leq 0.05$). Despite the country's smaller population, naming conventions lead to a higher frequency of homonyms. Nevertheless, the *LGBMRanker* model achieves strong performance, with mean values above 0.87 for MRR, 0.80 for Recall@1, and 0.97 for Recall@5, indicating that the available features are effective in distinguishing candidates even in the presence of many homonyms. Feature importance analysis shows that address-related features contribute 59.20% of the total importance, while age and occupation also provide meaningful signals, with contributions of 18.90% and 20.78%, respectively, reflecting a more balanced use of features compared to other public datasets.

IPUMS United States. *LGBMRanker* and *XGBRanker* achieve nearly identical performance across metrics, with no statistically significant differences ($W = 14$, $p > 0.05$). This dataset yields the highest performance among all evaluated datasets, with mean values above 0.99 for MRR, 0.98 for Recall@1, and 0.99 for Recall@5 for *LGBMRanker*. This near-perfect performance is primarily due to a limitation: the dataset represents only a 5% sample of the original census, which substantially reduces the likelihood of homonyms residing in the same household or neighborhood. As a result, address-related features become discriminative, accounting for an average of 84.52% of the total importance and effectively acting as a near-unique identifier. Therefore, the reported performance should be interpreted in the context of the available 5% sample, as a higher prevalence of homonyms in the full census population could reduce the discriminative power of address-related features.

Wikidata. *LGBMRanker* outperforms *XGBRanker* in MRR and Recall@1 ($W = 15$, $p \leq 0.05$), while no significant difference is observed for Recall@5 ($W = 10$, $p > 0.05$). The *LGBMRanker* model achieves strong performance, with mean values above 0.92 for MRR, 0.88 for Recall@1, and 0.98 for Recall@5. This dataset consists of public figures distributed globally, resulting in a low prevalence of homonyms. Consequently, age-related features dominate the ranking process, accounting for more than 66% of the total importance. This indicates that temporal information alone is often sufficient to identify the correct entity in this setting.

5.5.1. Impact of Candidate Set Size

Given that the *Brazilian Civil Registry* represents the most complex and realistic scenario, we further analyze model performance as a function of the candidate set size $|H_r|$. This analysis aims to assess whether simpler methods could suffice when the number of candidates is small, and how performance degrades as ambiguity increases.

Figure 3 (A) presents MRR, Recall@1, and Recall@5 as a function of $|H_r|$ in the interval $[2, 20]$. This controlled setting allows us to evaluate performance under relatively low ambiguity, where one might expect simpler methods to be competitive. However, the results show that both learning-to-rank models consistently outperform the baselines even for small candidate sets. This indicates that, even in low-ambiguity scenarios, the proposed approach is able to better exploit subtle signals and achieve more accurate rankings than heuristic or classification-based methods.

Figure 3 (B) extends this analysis to the full range of candidate set sizes, without

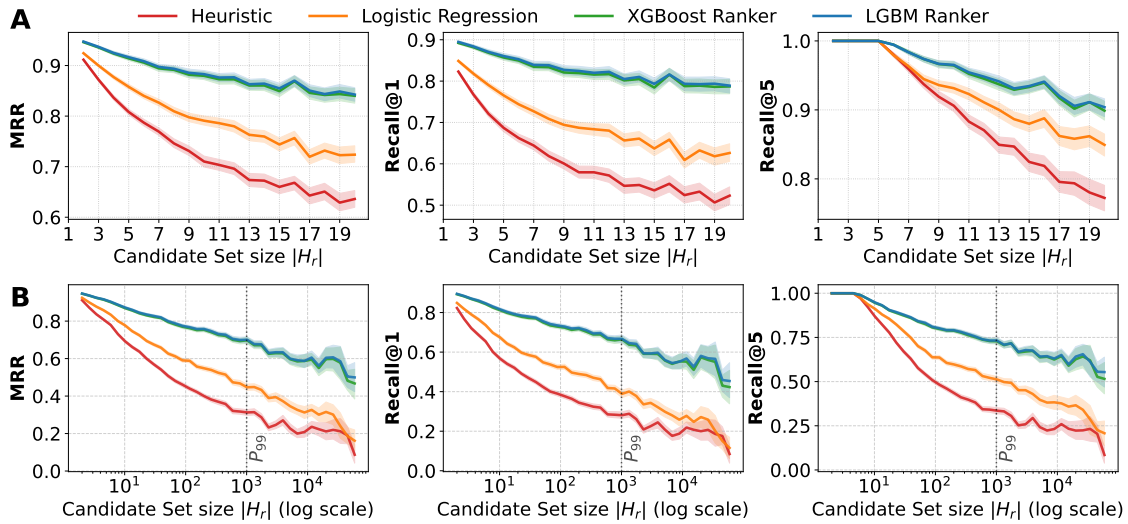

Figure 3. Metrics as a function of $|H_r|$ in the Brazilian Civil Registry. (A) $|H_r| \in [2, 20]$. (B) Full range of $|H_r|$ with a logarithmic x-axis.

imposing an upper bound on $|H_r|$. Across the entire range of candidate sizes, the proposed learning-to-rank models consistently outperform the baselines on all evaluation metrics. While exceptionally large candidate sets present a greater challenge, naturally softening the metrics in extreme scenarios, performance remains strong even at the 99th percentile for the best-performing model, *LGBMRanker*, with MRR reaching 0.70, Recall@1 0.66, and Recall@5 0.73. This suggests that the model maintains robust effectiveness in over 99% of the cases, despite the increased difficulty of the task.

6. Conclusion

In this work, we formalized the problem of linking relatives without identifiers in civil registry data as a Known-Item Search task and addressed it using a supervised learning-to-rank framework. Experimental results show that learning-to-rank models consistently outperform heuristic and classification-based baselines, achieving Recall@5 above 0.90 across five datasets, including a large-scale real-world Brazilian civil registry. The approach remains effective across varying levels of ambiguity, demonstrating robustness in both controlled and highly challenging settings. These results highlight the importance of modeling entity linking as a ranking problem and support the proposed method as a practical solution for real-world applications to better support human-in-the-loop validation.

Limitations and Future Work. This study assumes exact name matching during candidate generation, which may miss relatives due to typographical errors, spelling variations, or name changes after marriage. Future work should explore fuzzy name matching, which could increase the number of plausible candidates and shift the problem beyond the current known-item search setting.

Declaration of Generative AI use. Only for linguistic and grammatical refinement.

Acknowledgments. This work was supported by the Public Prosecution Service of the State of Minas Gerais (*Ministério Público do Estado de Minas Gerais* – MPMG), and by CNPq, CAPES, and FAPEMIG.

References

Albuquerque, D. L., Santos, V., Nack, P., et al. (2025). Language models are not a panacea: Combining them with domain knowledge and efficient indexes for entity linking. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, Porto Alegre, RS, Brasil. SBC.

Ali Omar, Z., Zamzuri, Z. H., Mohd Ariff, N., and Abu Bakar, M. A. (2023). Training data selection for record linkage classification. *Symmetry*, 15(5).

Archive, T. D., Centre, N. H. D., and Center, M. P. (2008). National Sample of the 1900 Census of Norway, Version 2.0.

Arora, A. and Dell, M. (2024). LinkTransformer: A unified package for record linkage with transformer language models. In Cao, Y., Feng, Y., and Xiong, D., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 221–231, Bangkok, Thailand. Association for Computational Linguistics.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer.

Garðarsdóttir, (n.d.). 1729 Census of Iceland, Version 1.0.

Ke, G., Meng, Q., Finley, T., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Marone, M., Weller, O., Fleshman, W., Yang, E., Lawrie, D., and Durme, B. V. (2025). mmbert: A modern multilingual encoder with annealed language learning.

Mon Myint, K. S. and Naing, W. W. (2023). Historical census data linkage: Graph-based household matching method. In *2023 IEEE Conference on Computer Applications (ICCA)*, pages 351–356.

Mrini, K., Nie, S., Gu, J., et al. (2022). Detection, disambiguation, re-ranking: Autoregressive entity linking as a multi-task problem. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1972–1983, Dublin, Ireland. Association for Computational Linguistics.

Nanayakkara, C. and Christen, P. (2022). Locality sensitive hashing with temporal and spatial constraints for efficient population record linkage. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 4354–4358, New York, NY, USA. Association for Computing Machinery.

Nations, U. (2014). *Principles and Recommendations for a Vital Statistics System, Revision 3*. United Nations.

Ornstein, J. T. (2025). Probabilistic record linkage using pretrained text embeddings. *Political Analysis*, page 1–12.

Price, J., Buckles, K., Van Leeuwen, J., and Riley, I. (2021). Combining family history and machine learning to link historical records: The census tree data set. *Explorations in Economic History*, 80:101391.

Ruggles, S., Cleveland, L. L., Lovatón Dávila, R., et al. (2025a). IPUMS International: Version 7.7 [dataset].

Ruggles, S., Flood, S., Sobek, M., et al. (2025b). IPUMS USA: Version 16.0 [dataset].

Santos, M. and Nascimento, D. (2023). Avaliando fatores de influência sobre algoritmos de aprendizado de máquina na etapa de classificação da resolução de entidades. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 63–75, Porto Alegre, RS, Brasil. SBC.

Scharpf, P., Breitinger, C., Spitz, A., et al. (2026). Entity linking with wikidata: A systematic literature review. *ACM Comput. Surv.*, 58(9).

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Shen, W., Wang, J., and Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Song, J., Nanayakkara, C., and Christen, P. (2025). Privately evaluating sensitive population record linkage without ground truth data. *International Journal of Data Science and Analytics*, 20:2971–2986.

Tian, S., Chen, Q., Comeau, D. C., et al. (2024). Pubmed computed authors in 2024: an open resource of disambiguated author names in biomedical literature. *Bioinformatics*, 40(11).

Vrandecic, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Wu, L., Petroni, F., Josifoski, M., et al. (2020). Scalable zero-shot entity linking with dense entity retrieval. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., et al. (2019). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhou, Q., Chen, W., Zhao, P.-P., et al. (2024). Towards effective author name disambiguation by hybrid attention. *Journal of Computer Science and Technology*, 39(4):929–950.

Ítalo Pereira and Ferreira, A. (2024). E-bela: Enhanced embedding-based entity linking approach. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web*, pages 115–123, Porto Alegre, RS, Brasil. SBC.