Kim Jinseok (Orcid ID: 0000-0001-6481-2065)

Title: Ethnicity-Based Name Partitioning for Author Name Disambiguation Using Supervised Machine Learning

Authors: Jinseok Kim, Jenna Kim, and Jason Owen-Smith

1. Jinseok Kim

Institute for Research on Innovation & Science, Survey Research Center, Institute for Social Research
University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104
jinseokk@umich.edu
ORCID id: 0000-0001-6481-2065

2. Jenna Kim

School of Information Sciences
University of Illinois at Urbana-Champaign
501 E. Daniel Street, Champaign, IL U.S.A. 61820
jkim682@illinois.edu

ORCID id: 0000-0001-7438-448X

3. Jason Owen-Smith

Department of Sociology, Institute for Social Research, University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104-2910
734-936-0463
jdos@umich.edu
ORCID id: 0000-0001-8007-1121

[Corresponding Author Information]

Jinseok Kim

Institute for Research on Innovation & Science, Survey Research Center, Institute for Social Research
University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104-2910
734-763-4994
jinseokk@umich.edu

[Financial Support Information]

This paper was supported by grants from the National Science Foundation (#1917663, #1760609, and #1561687).

Abstract

In several author name disambiguation studies, some ethnic name groups such as East Asian names are reported to be more difficult to disambiguate than others. This implies that disambiguation approaches might be improved if ethnic name groups are distinguished before disambiguation. We explore the potential of ethnic name partitioning by comparing performance of four machine learning algorithms trained and tested on the entire data or specifically on individual name groups. Results show that ethnicity-based name partitioning can substantially improve disambiguation performance because the individual models are better suited for their respective name group. The improvements occur across all ethnic name groups with different magnitudes. Performance gains in predicting matched name pairs outweigh losses in predicting nonmatched pairs. Feature (e.g., coauthor name) similarities of name pairs vary across ethnic name groups. Such differences may enable the development of ethnicity specific feature weights to improve prediction for specific ethic name categories. These findings are observed for three labeled data with a natural distribution of problem sizes as well as one in which all ethnic name groups are controlled for the same sizes of ambiguous names. This study is expected to motive scholars to group author names based on ethnicity prior to disambiguation.

Keywords: supervised machine learning; author name disambiguation; feature engineering, name ethnicity; data partition

Introduction and Background

A big challenge in managing digital libraries is that author names in bibliographic data are ambiguous because many authors have the same names (homonyms) or variant names are recorded for the same authors (synonyms). One study estimates that about two-thirds of author names in PubMed, the largest biomedicine digital library, are vulnerable to either or both of these two ambiguity types (Torvik & Smalheiser, 2009). Research findings obtained by mining bibliographic data can be distorted by merged and/or split author identities due to incorrect disambiguation (Fegley & Torvik, 2013; J. Kim & Diesner, 2015, 2016; Schulz, 2016). In addition, digital library users query author names most frequently (Islamaj Dogan, Murray, Névéol, & Lu, 2009). This means that the users will receive inaccurate information about research production, citation, and collaboration for authors if author name ambiguity is not properly resolved (Harzing, 2015; Strotmann & Zhao, 2012).

To address the challenge, researchers have proposed a variety of author name disambiguation (AND, hereafter) methods. Some scholars have used heuristics such as string-based matching (e.g., names that have the same full surname and forename initials are assumed to represent the same author), which is the most widely used approach in bibliometrics (Milojević, 2013). Others have developed rule-based programming and supervised/unsupervised machine learning techniques, as systemically reviewed in several papers (Ferreira, Gonçalves, & Laender, 2012; Hussain & Asghar, 2017; Sanyal, Bhowmick, & Das, 2019; Smalheiser & Torvik, 2009). In industry, several bibliographic data providers such as DBLP, Scopus, and Web of Science have disambiguated author names to improve their service quality (Kawashima & Tomizawa, 2015; J. Kim, 2018; Ley, 2009; Zhao, Rollins, Bai, & Rosen, 2017), while others still rely on the name string matching to output author-related search results.

Despite the differences in methods and datasets, a few AND studies have observed that some ethnic name groups (ENG, hereafter) (e.g., Chinese names) are more difficult to disambiguate than others (Deville et

al., 2014; J. Kim & Diesner, 2016; Strotmann & Zhao, 2012; Torvik & Smalheiser, 2009; J. Wu & Ding, 2013). This implies that author names may be better disambiguated if their associated ethnicities are considered as inputs in disambiguation models. But this possibility has been little explored. First, the observations made in several studies that certain ethnic names are harder to disambiguate are based on post-hoc evaluations of AND results. In other words, many of those studies did not integrate ethnic name partitions during machine learning. A very small number of studies have divided names into subgroups in their disambiguation model building (Chin et al., 2013; Louppe, Al-Natsheh, Susik, & Maguire, 2016) and evaluation process (Lerchenmueller & Sorenson, 2016). But their ethnic name categories are limited in number (e.g., dichotomy of Chinese vs non-Chinese; Caucasian, Asian, and Hispanic) or mixed up with racial distinctions based on the U.S. Social Security information (e.g., White, Black, Hispanic, Asian, etc.). Such racial classifications can be inappropriate for bibliographic data in which author names come from diverse regions around the world. In addition, those studies have typically used a single labeled data source, which makes it hard to expand and generalize their findings to other AND scenarios.

This study aims to empirically evaluate the effect ethnic name partitioning has on AND. In this study, author name disambiguation is a task to assign either ‘match’ or ‘nonmatch’ label to a pair of author name instances. For this, specifically, name instances are grouped into a block that share the same first forename initial and full surname and pairwisely compared with the block for their similarities over a set of features (e.g., coauthor name) to produce similarity scores. Machine learning algorithms combine the scores to learn weights of each feature to decide if a given pair of instances to refer to the same author or not. Although our work is motivated by the studies reviewed above and follows their common data pre-processing, blocking and machine learning steps, this paper differs from them in three important ways. First, this study evaluates AND performance by four different machine learning algorithms applied to four different labeled datasets before and after inclusion of a standard ethnic name group partition. Here, a name instance is assigned to an ethnic name group based on a name ethnicity classification system, *Ethnea*. Second, unlike traditional labeled data in which a specific ENG (i.e., Chinese) dominates, this study disambiguates new labeled data in which all ENGs are controlled to have the same numbers of instances, to demonstrate that performance changes induced by ethnic name partitioning may not be solely due to the well-known relationship between the number of cases and their ambiguity (more names, more ambiguity). Third, this study shows that different combinations of features (e.g., coauthor name and title words) appear to be related to AND performance for different ENGs suggesting future directions to further improve AND performance with ambiguous ethnic group names. The findings of this study can provide practical insights to researchers and practitioners who handle authority control in digital libraries. In the following sections, details on labeled data and setups for machine learning are described.

Method

Labeled Data and Pre-Processing

To measure the effect of ethnic name partition on machine learning for AND, this study disambiguates names in four labeled datasets – KISTI, AMINER, GESIS, and UM-IRIS. The first three datasets have been used in many AND studies to train and test machine learning algorithms (Cota, Ferreira, Nascimento, Gonçalves, & Laender, 2010; Ferreira, Veloso, Gonçalves, & Laender, 2014; Hussain & Asghar, 2018; J. Kim & Kim, 2018, In print; Momeni & Mayr, 2016; Alan Filipe Santana, Gonçalves, Laender, & Ferreira, 2017; Shin, Kim, Choi, & Kim, 2014; H. Wu, Li, Pei, & He, 2014; Zhu et al., 2018).

The last one is added to investigate how the ethnic name partition affects AND under the condition in which all ENGs are constrained to have the same numberss of ambigous name instances[1].

KISTI: Scientists at the Korea Institute of Science & Technology Information (KISTI) and Kyungsung University in Korea constructed this labeled dataset. It is made up of 41,673 author name instances that belong to 6,921 unique authors (Kang, Kim, Lee, Jung, & You, 2011)[2].

AMINER: Researchers in China and U.S. collaborated to create this labeled data to build and evaluate AND models for a computer science digital library, AMiner (Tang et al., 2008; X. Wang, Tang, Cheng, & Yu, 2011)[3]. It consists of 7,528 author name instances that refer to 1,546 unique authors.

GESIS: Scholars at the Leibniz Institute for the Social Sciences (GESIS) in Germany produced this labeled data. It contains author name instances of 5,408 unique authors (Momeni & Mayr, 2016) [4]. This study reuses the ‘Evaluation Set’ (29,965 author name instances of 2,580 unique authors) but with a few enhancements (J. Kim & Kim, 2020). Each author name instance is converted into the ‘surname, forename’ format and, through linking GESIS to its base DBLP data, is associated with the title of the paper in which it appears and the name of the conference or journal where the paper is published.

UM-IRIS: This dataset was generated by the researchers at the University of Michigan Institute for Research on Innovation & Science (UM-IRIS) and the University of Illinois through matching selected name instances in publication records to an authority database, ORCID (Kim & Owen-Smith, in print). First, author full names (e.g., ‘Brown, Michael’) that appear 50 times or more in MEDLINE-indexed publications published between 2000 and 2019 were listed[5]. Then, all instances of each selected full name (e.g., 158 instances of ‘Brown, Michael’ in MEDLINE) and their associated publication metadata were compared to 6 million researcher profiles in ORCID[6]. If an instance had a single match in the publication list of an ORCID researcher profile (matching on full name, paper title, and publication venue), the matched researcher’s ORCID id was assigned as an author label to the instance. Next, among the ORCID id-linked instances, those whose full names are associated with 5 or more ORCID ids (e.g., 6 unique ORCID researchers share the name ‘Brown, Michael’ which appear 158 times in MEDLINE) were randomly selected to produce 1,000 name instances for each of six ENGs. The resulting data contain 6,000 instances of 822 authors.

Four features – author name, coauthor name(s), paper title, and publication venue - are used as machine learning features because they have been widely used in algorithmic AND studies (Schulz, 2016; Song, Kim, & Kim, 2015) and are commonly available in the four labeled datasets. The string of each feature is stripped of non-alphabetical characters, converted into ASCII format, and lowercased. For title words, common English words like ‘the’ and ‘to’ are removed (i.e., stop-word listed) using the dictionary in Stanford NLP[7] and stemmed (e.g., ‘solution’ → ‘solut’) using the Porter’s algorithm[8]. Name instances in KISTI are converted into the full surname and first forename initial format (‘Wang, Wei’ → ‘Wang, W’) to make them more ambiguous (see J. Kim & Kim, 2020).

[1] This new labeled dataset was created following the idea of a reviewer who suggested that the impact of ethnic name partition on AND may be confounded by the size differences of ambiguous names.

[2] http://www.lbd.dcc.ufmg.br/lbd/collections/disambiguation/DBLP.tar.gz/at_download/file

[3] http://arnetminer.org/lab-datasets/disambiguation/rich-author-disambiguation-data.zip

[4] http://dx.doi.org/10.7802/1234

[5] https://www.nlm.nih.gov/bsd/medline.html

[6] https://orcid.org/

[7] https://github.com/stanfordnlp/CoreNLP/blob/master/data/edu/stanford/nlp/patterns/surface/stopwords.txt

[8] https://tartarus.org/martin/PorterStemmer/

Ethnic Name Group (ENG) Tagging

This study assigns an ENG tag to a name instance in each labeled dataset using the author name ethnicity classification database, *Ethnea*, developed by Torvik and Agarwal (2016)[9]. *Ethnea* is a collection of more than 9 million author name instances that are tagged one of 26 ENG classes based on the name's association with national-level geo-locations[10]. For example, "Wang, Wei" is classified as 'Chinese' as it is most frequently associated with organizations in China. However, *Ethnea* makes no distinctions based on any anthropological, cultural, or linguistic characteristics of authors. Instead it relies entirely on observations of names and geo-locations of their frequently associated institutions so an author named 'Wang, Wei' who was born in the U.S. and has never visited China would still be assigned a 'Chinese name' tag. We link all four labelled datasets to *Ethnea* and, if a matched name is found, assign that name's ENG tag to all its observed instances. If a queried name does not have a match in *Ethnea*, we search again using only the surname and assign the modal *Ethnea* ENG tag associated with it to all its instances. Table 1 summarizes the frequencies and ratios of ENG tags assigned by *Ethnea* to author name instances in each labeled dataset.

Table 1 shows the list of ENGs in each labeled data. Small-sized ENGs are excluded from analysis because most name instances in those ENGs tend to belong to a single author while a few instances referring to other author(s). When randomly split into training and test subsets for machine learning, these instances do not produce negative pairs at all.

[Table 1]

Chinese names represent the majority of ENGs in three labeled data. This is because these labeled data were created from computer science papers where Chinese researchers are particularly large contributors. In addition, as the three datasets were designed to collate challenging names to disambiguate, Chinese names that tend to be more ambiguous than other ENGs were over-sampled (Müller, Reitz, & Roy, 2017). In contrast, 6,000 instances in UM-IRIS are evenly distributed over six ENGs. For validation and reuse, these labeled data with ENG tags are publicly available[11]. Note that the original KISTI contains 41,673 name instances, whereas the ENG-tagged KISTI has 41,605 instances. Such discrepancy occurs because this paper uses the revised version of KISTI that corrects record errors and duplicates in the original data (J. Kim, 2018).

Machine Learning Process

Machine learning methods for AND can be divided into two groups: author assignment and author grouping (Ferreira et al., 2012). While the former aims to assign an author name instance to one of pre-disambiguated author name clusters, the latter aims to group all and only instances that belong to the same authors. This study takes the latter approach in evaluating the effect of ENG on AND. Specifically, author

[9] https://databank.illinois.edu/datasets/IDB-9087546

[10] 26 ethnicities include: African, Arab, Baltic, Caribbean, Chinese, Dutch, English, French, German, Greek, Hispanic, Hungarian, Indian, Indonesian, Israeli, Italian, Japanese, Korean, Mongolian, Nordic, Polynesian, Romanian, Slav, Thai, Turkish, and Vietnamese. In *Ethnea*, some name instances are assigned two ethnicities (e.g., "Jane Kim" → Korean-English) if the surname and forename of an author name are associated frequently with different ethnicities.

[11] Download link TBA

name instances in each labeled dataset are pairwise compared to assess whether a given instance pair of plausibly represents the same author (a match) or not (a nonmatch). Although some scholars take a further step to cluster pairwise comparisons (e.g., J. Kim & Kim, 2018; Levin, Krawczyk, Bethard, & Jurafsky, 2012; Louppe et al., 2016; Alan Filipe Santana et al., 2017), this study only evaluates disambiguation performance at a pair level (i.e., classification), following the practice of previous AND studies (e.g., Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004; Song et al., 2015; Treeratpituk & Giles, 2009; Vishnyakova, Rodriguez-Esteban, Ozol, & Rinaldi, 2016).

As the first machine learning step, author name instances in each labeled dataset are randomly divided into training (50%) and test (50%) subsets. Then, instances in each subset are put into blocks in which all member instances share the same full surname and first forename initial (e.g., 'Wang, W'). Only instances in the same block are compared for disambiguation. This blocking is typical in AND studies because it reduces computational complexity with only slight performance degradation (K. Kim, Sefid, & Giles, 2017; Torvik & Smalheiser, 2009). Next, instance pairs in the same block are compared to establish their similarity over four other data features: author name, coauthor name(s), paper title, and publication venue. To quantify how much a pair is similar over a feature, this study calculates the cosine similarity of Term (*n*-gram) Frequency for each feature (Han, Zha, & Giles, 2005; J. Kim & Kim, In print; Levin et al., 2012; Louppe et al., 2016; A. F. Santana, Gonçalves, Laender, & Ferreira, 2015; Treeratpituk & Giles, 2009). Specifically, the string of a feature is converted into an array of 2~4-grams (e.g., author name 'Wang, Wei' → 'wa|an|ng|gw|we|ei|wan|ang|ngw|gwe|wei|wang|angw|ngwe|gwei'). After the conversion, two *n*-gram arrays of an instance pair are compared to produce a cosine similarity score for the feature.

Besides the four basic features, ENGs are used as a feature set for ENG-aware disambiguation. For this, especially, an instance pair's ENG is encoded into a binary value (i.e., one-hot encoding) for a pre-defined set of ENGs[12]. For example, in AMINER, a pair of name instances ('Wang, Wei' and 'Wang, W.') is assigned either 'Yes' or 'No' for each of five ethnicities – Chinese ('Yes'), English ('No'), Indian ('No'), German ('No'), and Hispanic ('No') – as shown in Table 1. Table 2 shows examples of the cosine similarity scores calculated over four features and ENG encoding results for instance pairs.

[Table 2]

We focus on four algorithms – Gradient Boosting, Logistic Regression, Naïve Bayes, and Random Forest – for supervised machine learning that have been widely used as baselines or best performing methods in AND studies (e.g., Han et al., 2004; J. Kim & Kim, In print; K. Kim, Sefid, Weinberg, & Giles, 2018; Louppe et al., 2016; Song et al., 2015; Torvik & Smalheiser, 2009; Treeratpituk & Giles, 2009; Vishnyakova et al., 2016; J. Wang et al., 2012). In the first scenario, they are trained on the list of similarity scores and labels, as shown in Table 2 to learn relative weights for features and an absolute weight or threshold for instance pairs to be disambiguated without considering ENGs (→ ENG-ignorant learning). In the second scenario, the same algorithms are trained on the list of similarity scores, ENGs, and labels (→ ENG-aware disambiguation). Here, the ethnic name partition adds more features (dimensions) to each instance pair's feature set, allowing algorithms to combine the similarities of the expanded features. The machine learning procedure is implemented using the python Scikit-learn package. For Gradient Boosting, 500 estimators are used with max depth=9 and learning rate = 0.125. For

[12] As only instances in the same block in which they share at least the same full name and first forename initial are compared, all the pairs in the block have the same ethnicity tag.

Logistic Regression, L2 Regularization with class weight = 1 is used. Gaussian Naïve Bayes with maximum likelihood estimator is used for Naïve Bayes. For Random Forest, 500 trees are used after a grid search.

Trained algorithmic models are applied to the instance pairs in test subsets in which the cosine similarity is calculated for the four basic features and, in the second scenario, ethnicities are encoded in the same fashion but explicitly include ENG information. As in Table 2, an algorithmic model receives a set of feature similarity scores and, if ENG-aware disambiguation is conducted, a list of encoded ENGs for an instance pair to output a binary classification decision (match or nonmatch). Once trained, each algorithm produces a single score that predicts the probability of an instance pair being negative (nonmatch). If the predicted probability is above a certain threshold (> 0.5), the pair is decided to be a nonmatch, whereas if below the threshold, a match.

Performance Evaluation

We evaluate each algorithm's classification results on reserved test subsets of each labeled dataset by calculating precision and recall for positive (P; match) and negative (N; nonmatch) pairs respectively. In addition, we calculate the F1 score as a harmonic mean of precision and recall.

Specifically, precision for positive pairs (Prec-Pos) measures how many predicted match pairs are correct ones (true positives; TP) over the total number of predicted match pairs that may contain correct match pairs (true positives; TP) and incorrect match pairs (false positives; FP). In contrast, recall for positive pairs (Rec-Pos) measures the ratio of correct match pairs (true positives; TP) over the total number of true match pairs that may be predicted correctly as match pairs (true positives; TP) or incorrectly as nonmatch pairs (false negatives; FN).

$$Precision\ Positive\ (PrecPos) = \frac{Number\ of\ Correctly\ Predicted\ Match}{Number\ of\ Predicted\ Match} = \frac{TP}{(TP + FP)} \quad (1)$$

$$Recall\ Positive\ (RecPos) = \frac{Number\ of\ Correctly\ Predicted\ Match}{Number\ of\ True\ Match} = \frac{TP}{(TP + FN)} \quad (2)$$

$$F1\ Positive = \frac{2 \times PrecPos \times RecPos}{PrecPos\ +\ RecPos} \quad (3)$$

Likewise, precision for negative pairs (Prec-Neg) measures how many predicted nonmatch pairs are correct ones (true negatives; TN) over the total number of predicted nonmatch pairs that may contain correct nonmatch pairs (true negatives; TN) and incorrect nonmatch pairs (false negatives; FN). In contrast, recall for negative pairs (Rec-Neg) measures the ratio of correct nonmatch pairs (true negatives; TN) over the total number of true nonmatch pairs that may be predicted correctly as nonmatch pairs (true negatives; TN) or incorrectly as match pairs (false positives; FP).

$$Precision\ Negative\ (PrecNeg) = \frac{Number\ of\ Correctly\ Predicted\ Nonmatch}{Number\ of\ Predicted\ Nonmatch} = \frac{TN}{(TN + FN)} \quad (4)$$

$$Recall\ Negative\ (RecNeg) = \frac{Number\ of\ Correctly\ Predicted\ Nonmatch}{Number\ of\ True\ Nonmatch} = \frac{TN}{(TN + FP)} \quad (5)$$

$$F1\ Negative = \frac{2 \times PrecNeg \times RecNeg}{PrecNeg\ +\ RecNeg} \quad (6)$$

The metrics are calculated on the entire set of test results for each labeled dataset. We separately calculate performance measures for difference ENGs rather than averaging them across multiple ethnicity groups.

Results

Cross-Data Performance Evaluation

Figure 1 shows disambiguation results on KISTI, reporting precision and recall *before* and *after* ENG-aware disambiguation by four algorithms – Gradient Boosting (GB), Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF). Figure 1a shows that when ENGs are included as features, the algorithms tend to produce better precision in the prediction of positive (match) pairs than when they are not considered. This is shown by black bars ('After') being higher than stripped bars ('Before') in Figure 1a. This observation indicates that ethnic name partitioning helps algorithms increase the ratio of TP among predicted positive pairs (= TP + FP). This can be confirmed by checking the numbers of true and false positive pairs in Table 3. For example, when trained only on the four basic (non-ENG) features, LR predicts that 76,201 (= TP + FP = 55,998 + 20,203) pairs refer to the same authors (match) and 73.49% of the predictions are right (= TP/(TP + FP)). After trained on the same but ENG-tagged data, however, it predicts 170,432 pairs to be match sets, increasing its prediction accuracy this time to 77.08%.

[Figure 1]

[Table 3]

Ethnic name partitioning also reduces the number of falsely predicted nonmatch cases (FN), increasing recall in Figure 1b. Performance gains by ENG-aware disambiguation are more pronounced for recall than for precision, as evidenced by larger differences between 'Before' and 'After' bars for recall (Figure 1b) than those for precision (Figure 1a). In other words, ENG aware disambiguation across four common algorithms appears to reduce false negative predictions more than true positive predictions, potentially providing better performance for applications (such as network analysis) that are particularly sensitive to biases due to erroneous "lumping" of name instances that actually refer to different individuals. The improvements in precision and recall together increase the F1 scores by ENG-aware disambiguation (Figure 1c).

ENG-aware disambiguation also does a better job of accurately predicting non-match (negative pair) cases. The 'After' bars are taller than those of 'Before' in Figure 1d. Unlike the positive pair prediction in which ENG-aware disambiguation works in favor of both precision and recall by all algorithms, however, the performance gains in precision for negative pairs come with slightly decreased recall by GB and LR in Figure 1e. This means that while disambiguation models by GB and LR trained on ENG-added features are good at increasing the numbers of true nonmatch pairs among predicted nonmatch pairs (= TN +FN), they incorrectly predict that true nonmatch pairs match (FP predictions) more frequently than when they are trained on the four basic features alone. Reduced recall for negative pair prediction is, however, offset by increased precision, leading to the F1 scores by ENG-aware disambiguation being better than those by ENG-blind one in Figure 1f. Meanwhile, NB and RF still obtain improvements in both precision and recall as well as F1.

Algorithmic performances are also enhanced by ENG-aware disambiguation on AMINER, GESIS, and UM-IRIS. Figure 2 ~ 4 report that the algorithms trained on ENG-tagged data perform better than those trained only on the basic features across almost all metrics for both positive and negative pairs. NB models prove the exception, producing worse results in recall for positive pairs and in precision for negative pairs after ethnic name partition. However, this degraded performance is offset by increased precision for positive pairs and increased recall for negative pairs, respectively, so the overall performance metric (F1), which equally weights precision and recall, indicates an overall improvement due to the inclusion of ENG features.

[Figure 2]

[Figure 3]

[Figure 4]

Performance Evaluation per ENG

ENG-aware disambiguation produces substantial improvements in both precision and recall for predicting match and nonmatch instance pairs in different labeled datasets. But are those improvements uniform across different ENGs? If not, a more nuanced approach to model evaluation may be necessary. To answer this question, we compare performance changes due to ENG-aware disambiguation within ENG groups. For this, precision, recall, and F1 scores for positive and negative pairs predicted by four algorithms are calculated separately for instance pairs that belong to the same ENG in each of four labeled data: 4 algorithms × 4 data = 16 evaluations. Presenting all the results at the same time would consume too much space in this paper. So, we present random forest (RF) predictions on the GESIS dataset as an illustration for the purposes of this discussion. Reports of other algorithms and data are presented in a supplementary document attached to this paper.

Figure 5 shows the by ENG performance metrics for the RF algorithm trained on GESIS with and without ENG-aware disambiguation. The ENG-aware disambiguation leads to better precision (positive pairs; Figure 5A) and recall (negative pairs; Figure 5E) for Chinese names but worse precision (positive pairs) and recall (negative pairs) for other ethnicities. In contrast, name disambiguation for Chinese names results in lower recall (positive pairs; Figure 4B) and precision (negative pairs; Figure 4D) than those for other ENGs. Similar patterns are observed for other algorithms tested on GESIS (see Figure S5~S8 in Supplementary Material). This suggests that the effect of ENG-aware disambiguation occurs in different ways for different ENGs. Thus, its application can be beneficial in some instances but detrimental in others. Variations in the effects of ENG-aware disambiguation on precision and recall for positive and negative pair prediction across ethnicity groups suggest that care must be taken to design disambiguation strategies that fit particular analytic or empirical needs.

[Figure 5]

These observations can be explained as follows. ENGs have different distributions of similarity scores over the four basic (non-ethnicity) features we use. Figure 6 presents the feature similarity score distributions per ENG for positive (left) and negative (right) pairs in the GESIS test data. Training and test subsets show similar distributions in each labeled dataset. For visual simplicity, a score is rounded up into nearest bins with intervals of 0.1 on *x*-axis and the ratios of the numbers of scores in the same bin over all scores are plotted on *y*-axis. A solid red line represents the distribution of all instance pairs regardless of ENG.

In Figure 6, each ENG has different distributions of, for example, 'COAUTHOR' similarity scores for both positive and negative pairs (Figure 6C and 6D). So, the four algorithms come to use different 'coauthor' similarity score distributions in ENG-aware disambiguation. Such heterogeneous distributions also occur for other features but with different variations of differences. For example, 'VENUE' distributions in Figure 6G and 6H differ less across ENGs than do 'COAUTHOR' distributions. Because ENG-aware disambiguation allows training and testing on different feature similarity score distributions for each ENG, the algorithms combine features using different weightings for each ethnicity, producing different predictions for name pairs with the same feature similarity scores but different ENG tags. In other words, this method takes into account the likelihood that researchers in different ENGs organize their scientific work differently, favoring distinct co-authorship and publication venue patterns. This also occurs in disambiguation of other labeled data, whose feature similarity score distributions are reported in Figure S17 ~ S20 in Supplementary Material.

[Figure 6]

Figure 5 also shows that some ENGs manifest substantial improvements in recall for positive pairs (Figure 5B) but degraded recall for negative pairs (Figure 5E). This might be explained in two ways. In our 'before' (ENG-unaware) case, algorithms combine features to produce per-feature weights for positive pairs based on feature similarity scores aggregated across multiple ENGs that can have very different feature distributions. Such aggregated distributions cannot effectively capture the single match patterns specific to each ENG, which seem to lead models to falsely predict positive pairs as negative ones (FN), reducing the recall for positive pairs. Conversely, increased recall for positive pairs after ENG-aware disambiguation means that the algorithms trained and tested on ENG-tagged data successfully produce per-feature weights optimized to each ENG, thus making better predictions that push up the recall scores for many ENGs.

Second, decreased negative pair recall after ENG-aware disambiguation means that the algorithms trained and tested on ENG-tagged data fail to produce proper per-feature weights for accurately predicting nonmatch for known negative pairs. When the algorithms are trained only on the four basic (non-ethnicity) features, they do a better job of predicting nonmatch pairs based on aggregated feature similarity distributions that are invariant across particular ENGs. In other words, feature distributions aggregated across ENGs appear to be more effective for predicting negative case pairs while ENG-aware disambiguation techniques more accurately capture positive pairs.

These observations imply that disambiguation models for positive pair prediction would be improved by ENG-aware procedures, while nonmatch patterns for negative pair prediction can aggregate across ENGs (J. Kim & Kim, 2018). Table 4 shows that in the GESIS training data, each ENG has different sizes of positive and negative (pairwise) pairs. CHINESE name instances produce the largest numbers of positive

(≈ 146K) and negative pairs (≈ 551K), while ITALIAN name instances generate around a few thousand positive and a few hundred negative pairs. In other training data, CHINESE pairs constitute substantially large proportions (KISTI: 71.28 % and AMINER: 91.26 %) or over one-third (UM-IRIS: 37.54 %) of all negative pairs, while other ENG pairs make up small or less-than-expected (approximately 17% per ENG in UM-IRIS) proportions. In contrast, the numbers of positive pairs are less concentrated (GESIS, KISTI, and AMINER) or more evenly distributed (UM-IRIS) for positive pairs than those for negative pairs.

[Table 4]

[Table 5]

[Table 6]

[Table 7]

As noted above for Figure 5 and observed in other labeled data (see Figure S1 ~ S16 in Supplementary Material), the algorithms work better in finding more true negative pairs even for non-CHINESE name pairs when they are trained on data in which ethnic name partitioning is not performed (‘Before’) and, thus, negative pairs are dominated by CHINESE ones as shown in Table 4~6. This implies that the nonmatch patterns in CHINESE name pairs are applicable to predicting nonmatch pairs for other ENGs. In contrast, during ENG-aware disambiguation, the algorithms come to rely on the small-size negative pairs that may skew or distort true nonmatch patterns for some ENGs. This seems to result in the decreased recall in predicting negative pairs (i.e., many true negatives classified as false positives, which reduces precision for positive pair prediction), while increasing slightly precision in predicting negative pairs.

Despite the aforementioned conflicting changes in precision and recall per ENG, the overall performance by the four algorithms on the whole test set are shown in Figure 1~4 to substantially increase across the four labeled data after ethnic name partitioning is included in machine learning. One reason would be that performance gains outweigh losses at each ENG level overall. Another reason would be that especially for KISTI, AMINER, and GESIS, the improved performances in disambiguating CHINESE that constitute the majority of name instances may affect the overall evaluation results. As shown by the case of UM-IRIS in which ENG sizes are controlled to be equal, however, the overall performance improvements can be observed for all the ENGs by ENG-aware disambiguation. As such, this study illustrates that the ethnic name partition can be truly effective in improving disambiguation performances.

Discussion

These results suggest that AND tasks may produce better results by using ethnic name partition in machine learning. Considering that adding more features can improve generally machine learning performances, the enhanced disambiguation performances by ENG partitioning might not be a surprise. With that said, the real contribution of this study would be that it demonstrates many machine learning based disambiguation models have a potential to be improved by introducing ethnic name grouping into ambiguous data without additional collection of feature information.

To fully realize this potential, however, a few issues need to be addressed. First, ENG tagging can be a non-trivial task that requires an intricate algorithmic technique itself. Thanks to the ENG classification system developed and publicly shared by Torvik and Agarwal (2016), this study could assign ENGs to the names in four labeled data. Although *Ethnea* was constructed based on more than 9 million author name instances in PubMed, the world largest biomedicine library, it is unknown how well it can help us tag ENGs to names in other fields. Ideally, *Ethnea* may be updated regularly to reflect new author names entering bibliographic data in various fields. Practically, further research may be focused on finding out a set of ENGs that are most influential in improving disambiguation results and thus simplifying ENG tagging for author name disambiguation (e.g., CHINESE vs Non-CHINESE).

Second, the findings of this study were based on three labeled data (KISTI, AMINER, and GESIS) in which CHINESE names are dominant and the overall performance improvements were heavily affected by those for CHINESE name instances. To overcome such an imbalance of instance distribution in labeled data, a new labeled data (UM-IRIS) were created in a way that six ENGs have the same amount of ambiguous name instances. Disambiguation results from the new labeled data were in line with those from other three labeled data. In addition, all ENGs including CHINESE were able to obtain gains in disambiguation performances. But all these findings were obtained from small-sized labeled data, whether they are biased or controlled for ENG sizes. So, it is still unknown whether such improvements are achievable in author name disambiguation for large-scale bibliographic data in which ENG composition may be quite different from those in the labeled data used in this paper.

Another issue would be that there can be other features than the four used in this study that can lead ethnic name partition to different AND performances. For example, English authors may appear in publication records that are more complete in affiliation information and use more diverse title terms. Meanwhile, Chinese authors may tend to work with coauthors who have similar names in same institutions. Various features need to be explored to study further the impact of ethnic name partition on AND.

Fourth, ENG-aware disambiguation may be beneficial for positive pair prediction but not so much for negative pair prediction. This was illustrated in Figure 5 above and Figure S1 ~ S16 in Supplementary Material by the dramatically decreased recall in negative pair predictions for many ENGs. It was contrasted with the substantial increase of precision in positive pair prediction for those ENGs. This study speculates that by ethnic name partitioning, classifiers become stricter for CHINESE pairs while relaxed for other ENG pairs. In other words, a pairs of CHINESE instances that would be classified as 'match' before partitioning are classified as 'nonmatch' after partitioning (PREC-POS↑; REC-POS↓), while 'nonmatch' pairs of other ENG instances as matched ones (PREC-POS↓; REC-POS↑). This might be because while some CHINESE pairs sharing coauthor names, venue names, or title words refer to different authors, other ethnic names sharing the features are more likely to represent the same authors (see Figure 6 B, D, F, and H in which Chinese name pair share is denoted in square). During training, such different similarity patterns are mixed up before partitioning but distinguished after it. Another conjecture is that due to the relaxed classification after partitioning, true negative pairs of other-than-CHINESE ENGs are falsely classified as false positive pairs (PREC-POS↓; REC-NEG↓). As the sizes of negative pairs in most non-CHINESE ENGs are smaller than positive pairs (see Table 4~7), misclassified negative pairs have larger impacts on REC-NEG than on PREC-POS across the ENGs. But these conjectures are based on the observations on labeled data in which Chinese name instances are prevalent. Using an additional labeled data with controlled ENG sizes, however, the conjecture has been confirmed. But only six ENGs in a small dataset were considered for analysis. More ENGs need to be investigated to check if this conjecture holds good under the different combinations of ENGs.

Conclusion and Discussion

This study evaluated the effects ethnic name partitioning has on author name disambiguation (AND) using machine learning methods. For this, author name instances in four labeled datasets were disambiguated under two scenarios. First, similarity scores of instance pairs over four basic features – author name, coauthor names, paper title, and publication venue – were used to train and test disambiguation algorithms. Second, in addition to the basic features, ethnic name group (ENGs) were tagged to name instances to allow algorithms to build models that are optimized to each ENG. Comparisons of disambiguation performances before and after ENG-aware disambiguation showed that using ethnic name partition can substantially improve algorithmic performances. Such performance improvements occurred across all ENGs, although performance gains and losses at each ENG level were observed in different ways depending on the types of measures – precision or recall – and target classifications – positive (match) or negative (nonmatch) pairs.

As detailed in the discussion above, ethnic name partition requires further research for us to understand better its impact on author name disambiguation and apply it to disambiguation tasks for digital libraries that are struggling with authority control over fast-growing ambiguous author names. This study is expected to motivate scholars and practitioners to study toward that direction by demonstrating the potential of ENG-aware disambiguation in improving disambiguation performances.

Acknowledgements

This paper was supported by grants from the National Science Foundation (#1917663, #1760609, and #1561687). The authors would like to thank anonymous reviewers for their comments to improve the paper.

References

Chin, A. W.-S., Juan, Y.-C., Zhuang, Y., Wu, F., Tung, H.-Y., Yu, T., . . . et al. (2013). *Effective string processing and matching for author disambiguation*. Paper presented at the Proceedings of the 2013 KDD Cup 2013 Workshop, Chicago, Illinois. https://doi.org/10.1145/2517288.2517295

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. F. (2010). An Unsupervised Heuristic-Based Hierarchical Method for Name Disambiguation in Bibliographic Citations. *Journal of the American Society for Information Science and Technology, 61*(9), 1853-1870. doi:10.1002/asi.21363

Deville, P., Wang, D. S., Sinatra, R., Song, C. M., Blondel, V. D., & Barabási, A.-L. (2014). Career on the Move: Geography, Stratification, and Scientific Impact. *Scientific Reports, 4*. doi:10.1038/srep04770

Fegley, B. D., & Torvik, V. I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLOS ONE, 8*(7). doi:10.1371/journal.pone.0070299

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. *Sigmod Record, 41*(2), 15-26.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. F. (2014). Self-Training Author Name Disambiguation for Information Scarce Scenarios. *Journal of the Association for Information Science and Technology, 65*(6), 1257-1278. doi:10.1002/asi.22992

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). *Two Supervised Learning Approaches for Name Disambiguation in Author Citations.* Paper presented at the Jcdl 2004: Proceedings of the Fourth Acm/Ieee Joint Conference on Digital Libraries, Tucson, Arizona.

Han, H., Zha, H. Y., & Giles, C. L. (2005). Name disambiguation in author citations using a K-way spectral clustering method. *Proceedings of the 5th Acm/Ieee Joint Conference on Digital Libraries, Proceedings*, 334-343. doi:Doi 10.1145/1065385.1065462

Harzing, A. W. (2015). Health warning: might contain multiple personalities-the problem of homonyms in Thomson Reuters Essential Science Indicators. *Scientometrics, 105*(3), 2259-2270. doi:10.1007/s11192-015-1699-y

Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review, 32*, e22. doi:10.1017/S0269888917000182

Hussain, I., & Asghar, S. (2018). DISC: Disambiguating homonyms using graph structural clustering. *Journal of Information Science, 44*(6), 830-847. doi:10.1177/0165551518761011

Islamaj Dogan, R., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database : the journal of biological databases and curation, 2009*, bap018-bap018. doi:10.1093/database/bap018

Kang, I. S., Kim, P., Lee, S., Jung, H., & You, B. J. (2011). Construction of a Large-Scale Test Set for Author Disambiguation. *Information Processing & Management, 47*(3), 452-465. doi:10.1016/j.ipm.2010.10.001

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics, 103*(3), 1061-1071. doi:10.1007/s11192-015-1580-z

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: A case of DBLP. *Scientometrics, 116*(3), 1867-1886. doi:10.1007/s11192-018-2824-5

Kim, J., & Diesner, J. (2015). The effect of data pre-processing on understanding the evolution of collaboration networks. *Journal of Informetrics, 9*(1), 226-236. doi:10.1016/j.joi.2015.01.002

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology, 67*(6), 1446-1461. doi:10.1002/asi.23489

Kim, J., & Owen-Smith, J. (In print). ORCID-linked labeled data for evaluating author name disambiguation at scale. Scientometrics. doi:10.1007/s11192-020-03826-6

Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics, 117*(1), 511-526. doi:10.1007/s11192-018-2865-9

Kim, J., & Kim, J. (2020). Effect of Forename String on Author Name Disambiguation. *Journal of the Association for Information Science and Technology, 71*(7), 839-855. doi:10.1002/asi.24298

Kim, K., Sefid, A., & Giles, C. L. (2017). Scaling Author Name Disambiguation with CNF Blocking. *arXiv preprint arXiv:1709.09657*.

Kim, K., Sefid, A., Weinberg, B. A., & Giles, C. L. (2018). *A Web Service for Author Name Disambiguation in Scholarly Databases.* Paper presented at the 2018 IEEE International Conference on Web Services (ICWS).

Lerchenmueller, M. J., & Sorenson, O. (2016). Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists. *PLOS ONE, 11*(7), e0158731. doi:10.1371/journal.pone.0158731

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-Based Bootstrapping for Large-Scale Author Disambiguation. *Journal of the American Society for Information Science and Technology, 63*(5), 1030-1047. doi:10.1002/asi.22621

Ley, M. (2009). DBLP: some lessons learned. *Proceedings of the VLDB Endowment, 2*(2), 1493-1500.

Louppe, G., Al-Natsheh, H. T., Susik, M., & Maguire, E. J. (2016). ENG Sensitive Author Disambiguation Using Semi-supervised Learning. *Knowledge Engineering and Semantic Web, Kesw 2016, 649*, 272-287. doi:10.1007/978-3-319-45880-9_21

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics, 7*(4), 767-773. doi:10.1016/j.joi.2013.06.006

Momeni, F., & Mayr, P. (2016). *Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names.* Paper presented at the 20th international Conference on Theory and Practice of Digital Libraries (TPDL 2016), Hannover, Germany.

Müller, M. C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics, 111*(3), 1467-1500. doi:10.1007/s11192-017-2363-5

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2015). On the Combination of Domain-Specific Heuristics for Author Name Disambiguation: The Nearest Cluster Method. *International Journal on Digital Libraries, 16*(3-4), 229-246. doi:10.1007/s00799-015-0158-y

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2017). Incremental author name disambiguation by exploiting domain-specific heuristics. *Journal of the Association for Information Science and Technology, 68*(4), 931-945. doi:10.1002/asi.23726

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2019). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*. doi:10.1177/0165551519888605

Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics, 107*(3), 1283-1298. doi:10.1007/s11192-016-1892-7

Shin, D., Kim, T., Choi, J., & Kim, J. (2014). Author Name Disambiguation Using a Graph Model with Node Splitting and Merging Based on Bibliographic Information. *Scientometrics, 100*(1), 15-50. doi:10.1007/s11192-014-1289-4

Smalheiser, N. R., & Torvik, V. I. (2009). Author Name Disambiguation. *Annual Review of Information Science and Technology, 43*, 287-313.

Song, M., Kim, E. H. J., & Kim, H. J. (2015). Exploring Author Name Disambiguation on PubMed-Scale. *Journal of Informetrics, 9*(4), 924-941. doi:10.1016/j.joi.2015.08.004

Strotmann, A., & Zhao, D. Z. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology, 63*(9), 1820-1833. doi:10.1002/asi.22695

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). *ArnetMiner: extraction and mining of academic social networks*. Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA.

Torvik, V. I., & Agarwal, S. (2016). *Ethnea: An instance-based ENG classifier based on geo-coded author names in a large-scale bibliographic database*. Paper presented at the Library of Congress International Symposium on Science of Science, Washington D.C. http://hdl.handle.net/2142/88927

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *Acm Transactions on Knowledge Discovery from Data, 3*(3). doi:10.1145/1552303.1552304

Treeratpituk, P., & Giles, C. L. (2009). Disambiguating Authors in Academic Publications using Random Forests. *JCDL 2009: Proceedings of the 2009 Acm/Ieee Joint Conference on Digital Libraries*, 39-48.

Vishnyakova, D., Rodriguez-Esteban, R., Ozol, K., & Rinaldi, F. (2016, dec). *Author Name Disambiguation in MEDLINE Based on Journal Descriptors and Semantic Types.* Paper presented at the Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), Osaka, Japan.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics, 93*(2), 391-411. doi:10.1007/s11192-012-0681-1

Wang, X., Tang, J., Cheng, H., & Yu, P. S. (2011). *ADANA: Active Name Disambiguation*. Paper presented at the 2011 IEEE 11th International Conference on Data Mining.

Wu, H., Li, B., Pei, Y. J., & He, J. (2014). Unsupervised author disambiguation using Dempster-Shafer theory. *Scientometrics, 101*(3), 1955-1972. doi:10.1007/s11192-014-1283-x

Wu, J., & Ding, X. H. (2013). Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics, 96*(3), 683-697. doi:10.1007/s11192-013-0978-8

Zhao, Z., Rollins, J., Bai, L., & Rosen, G. (2017, Oct. 2017). *Incremental Author Name Disambiguation for Scientific Citation Data.* Paper presented at the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Zhu, J., Wu, X., Lin, X., Huang, C., Fung, G. P. C., & Tang, Y. (2018). A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering. *Scientometrics, 114*(3), 781-794. doi:10.1007/s11192-017-2611-8

Title: Ethnicity-Based Name Partitioning for Author Name Disambiguation Using Supervised Machine Learning

Authors: Jinseok Kim, Jenna Kim, and Jason Owen-Smith

1. Jinseok Kim

Institute for Research on Innovation & Science, Survey Research Center, Institute for Social Research
University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104
jinseokk@umich.edu
ORCID id: 0000-0001-6481-2065

2. Jenna Kim

School of Information Sciences
University of Illinois at Urbana-Champaign
501 E. Daniel Street, Champaign, IL U.S.A. 61820
jkim682@illinois.edu

ORCID id: 0000-0001-7438-448X

3. Jason Owen-Smith

Department of Sociology, Institute for Social Research, University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104-2910
734-936-0463
jdos@umich.edu
ORCID id: 0000-0001-8007-1121

[Corresponding Author Information]

Jinseok Kim

Institute for Research on Innovation & Science, Survey Research Center, Institute for Social Research
University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104-2910
734-763-4994
jinseokk@umich.edu

[Financial Support Information]

This paper was supported by grants from the National Science Foundation (#1917663, #1760609, and #1561687).

Abstract

In several author name disambiguation studies, some ethnic name groups such as East Asian names are reported to be more difficult to disambiguate than others. This implies that disambiguation approaches might be improved if ethnic name groups are distinguished before disambiguation. We explore the potential of ethnic name partitioning by comparing performance of four machine learning algorithms trained and tested on the entire data or specifically on individual name groups. Results show that ethnicity-based name partitioning can substantially improve disambiguation performance because the individual models are better suited for their respective name group. The improvements occur across all ethnic name groups with different magnitudes. Performance gains in predicting matched name pairs outweigh losses in predicting nonmatched pairs. Feature (e.g., coauthor name) similarities of name pairs vary across ethnic name groups. Such differences may enable the development of ethnicity specific feature weights to improve prediction for specific ethic name categories. These findings are observed for three labeled data with a natural distribution of problem sizes as well as one in which all ethnic name groups are controlled for the same sizes of ambiguous names. This study is expected to motive scholars to group author names based on ethnicity prior to disambiguation.

Keywords: supervised machine learning; author name disambiguation; feature engineering, name ethnicity; data partition

Introduction and Background

A big challenge in managing digital libraries is that author names in bibliographic data are ambiguous because many authors have the same names (homonyms) or variant names are recorded for the same authors (synonyms). One study estimates that about two-thirds of author names in PubMed, the largest biomedicine digital library, are vulnerable to either or both of these two ambiguity types (Torvik & Smalheiser, 2009). Research findings obtained by mining bibliographic data can be distorted by merged and/or split author identities due to incorrect disambiguation (Fegley & Torvik, 2013; J. Kim & Diesner, 2015, 2016; Schulz, 2016). In addition, digital library users query author names most frequently (Islamaj Dogan, Murray, Névéol, & Lu, 2009). This means that the users will receive inaccurate information about research production, citation, and collaboration for authors if author name ambiguity is not properly resolved (Harzing, 2015; Strotmann & Zhao, 2012).

To address the challenge, researchers have proposed a variety of author name disambiguation (AND, hereafter) methods. Some scholars have used heuristics such as string-based matching (e.g., names that have the same full surname and forename initials are assumed to represent the same author), which is the most widely used approach in bibliometrics (Milojević, 2013). Others have developed rule-based programming and supervised/unsupervised machine learning techniques, as systemically reviewed in several papers (Ferreira, Gonçalves, & Laender, 2012; Hussain & Asghar, 2017; Sanyal, Bhowmick, & Das, 2019; Smalheiser & Torvik, 2009). In industry, several bibliographic data providers such as DBLP, Scopus, and Web of Science have disambiguated author names to improve their service quality (Kawashima & Tomizawa, 2015; J. Kim, 2018; Ley, 2009; Zhao, Rollins, Bai, & Rosen, 2017), while others still rely on the name string matching to output author-related search results.

Despite the differences in methods and datasets, a few AND studies have observed that some ethnic name groups (ENG, hereafter) (e.g., Chinese names) are more difficult to disambiguate than others (Deville et

al., 2014; J. Kim & Diesner, 2016; Strotmann & Zhao, 2012; Torvik & Smalheiser, 2009; J. Wu & Ding, 2013). This implies that author names may be better disambiguated if their associated ethnicities are considered as inputs in disambiguation models. But this possibility has been little explored. First, the observations made in several studies that certain ethnic names are harder to disambiguate are based on post-hoc evaluations of AND results. In other words, many of those studies did not integrate ethnic name partitions during machine learning. A very small number of studies have divided names into subgroups in their disambiguation model building (Chin et al., 2013; Louppe, Al-Natsheh, Susik, & Maguire, 2016) and evaluation process (Lerchenmueller & Sorenson, 2016). But their ethnic name categories are limited in number (e.g., dichotomy of Chinese vs non-Chinese; Caucasian, Asian, and Hispanic) or mixed up with racial distinctions based on the U.S. Social Security information (e.g., White, Black, Hispanic, Asian, etc.). Such racial classifications can be inappropriate for bibliographic data in which author names come from diverse regions around the world. In addition, those studies have typically used a single labeled data source, which makes it hard to expand and generalize their findings to other AND scenarios.

This study aims to empirically evaluate the effect ethnic name partitioning has on AND. In this study, author name disambiguation is a task to assign either 'match' or 'nonmatch' label to a pair of author name instances. For this, specifically, name instances are grouped into a block that share the same first forename initial and full surname and pairwisely compared with the block for their similarities over a set of features (e.g., coauthor name) to produce similarity scores. Machine learning algorithms combine the scores to learn weights of each feature to decide if a given pair of instances to refer to the same author or not. Although our work is motivated by the studies reviewed above and follows their common data pre-processing, blocking and machine learning steps, this paper differs from them in three important ways. First, this study evaluates AND performance by four different machine learning algorithms applied to four different labeled datasets before and after inclusion of a standard ethnic name group partition. Here, a name instance is assigned to an ethnic name group based on a name ethnicity classification system, *Ethnea*. Second, unlike traditional labeled data in which a specific ENG (i.e., Chinese) dominates, this study disambiguates new labeled data in which all ENGs are controlled to have the same numbers of instances, to demonstrate that performance changes induced by ethnic name partitioning may not be solely due to the well-known relationship between the number of cases and their ambiguity (more names, more ambiguity). Third, this study shows that different combinations of features (e.g., coauthor name and title words) appear to be related to AND performance for different ENGs suggesting future directions to further improve AND performance with ambiguous ethnic group names. The findings of this study can provide practical insights to researchers and practitioners who handle authority control in digital libraries. In the following sections, details on labeled data and setups for machine learning are described.

Method

Labeled Data and Pre-Processing

To measure the effect of ethnic name partition on machine learning for AND, this study disambiguates names in four labeled datasets – KISTI, AMINER, GESIS, and UM-IRIS. The first three datasets have been used in many AND studies to train and test machine learning algorithms (Cota, Ferreira, Nascimento, Gonçalves, & Laender, 2010; Ferreira, Veloso, Gonçalves, & Laender, 2014; Hussain & Asghar, 2018; J. Kim & Kim, 2018, In print; Momeni & Mayr, 2016; Alan Filipe Santana, Gonçalves, Laender, & Ferreira, 2017; Shin, Kim, Choi, & Kim, 2014; H. Wu, Li, Pei, & He, 2014; Zhu et al., 2018).

The last one is added to investigate how the ethnic name partition affects AND under the condition in which all ENGs are constrained to have the same numberss of ambigous name instances[1].

KISTI: Scientists at the Korea Institute of Science & Technology Information (KISTI) and Kyungsung University in Korea constructed this labeled dataset. It is made up of 41,673 author name instances that belong to 6,921 unique authors (Kang, Kim, Lee, Jung, & You, 2011)[2].

AMINER: Researchers in China and U.S. collaborated to create this labeled data to build and evaluate AND models for a computer science digital library, AMiner (Tang et al., 2008; X. Wang, Tang, Cheng, & Yu, 2011)[3]. It consists of 7,528 author name instances that refer to 1,546 unique authors.

GESIS: Scholars at the Leibniz Institute for the Social Sciences (GESIS) in Germany produced this labeled data. It contains author name instances of 5,408 unique authors (Momeni & Mayr, 2016) [4]. This study reuses the ‘Evaluation Set’ (29,965 author name instances of 2,580 unique authors) but with a few enhancements (J. Kim & Kim, 2020). Each author name instance is converted into the ‘surname, forename’ format and, through linking GESIS to its base DBLP data, is associated with the title of the paper in which it appears and the name of the conference or journal where the paper is published.

UM-IRIS: This dataset was generated by the researchers at the University of Michigan Institute for Research on Innovation & Science (UM-IRIS) and the University of Illinois through matching selected name instances in publication records to an authority database, ORCID (Kim & Owen-Smith, in print). First, author full names (e.g., ‘Brown, Michael’) that appear 50 times or more in MEDLINE-indexed publications published between 2000 and 2019 were listed[5]. Then, all instances of each selected full name (e.g., 158 instances of ‘Brown, Michael’ in MEDLINE) and their associated publication metadata were compared to 6 million researcher profiles in ORCID[6]. If an instance had a single match in the publication list of an ORCID researcher profile (matching on full name, paper title, and publication venue), the matched researcher’s ORCID id was assigned as an author label to the instance. Next, among the ORCID id-linked instances, those whose full names are associated with 5 or more ORCID ids (e.g., 6 unique ORCID researchers share the name ‘Brown, Michael’ which appear 158 times in MEDLINE) were randomly selected to produce 1,000 name instances for each of six ENGs. The resulting data contain 6,000 instances of 822 authors.

Four features – author name, coauthor name(s), paper title, and publication venue - are used as machine learning features because they have been widely used in algorithmic AND studies (Schulz, 2016; Song, Kim, & Kim, 2015) and are commonly available in the four labeled datasets. The string of each feature is stripped of non-alphabetical characters, converted into ASCII format, and lowercased. For title words, common English words like ‘the’ and ‘to’ are removed (i.e., stop-word listed) using the dictionary in Stanford NLP[7] and stemmed (e.g., ‘solution’ → ‘solut’) using the Porter’s algorithm[8]. Name instances in KISTI are converted into the full surname and first forename initial format (‘Wang, Wei’ → ‘Wang, W’) to make them more ambiguous (see J. Kim & Kim, 2020).

[1] This new labeled dataset was created following the idea of a reviewer who suggested that the impact of ethnic name partition on AND may be confounded by the size differences of ambiguous names.

[2] http://www.lbd.dcc.ufmg.br/lbd/collections/disambiguation/DBLP.tar.gz/at_download/file

[3] http://arnetminer.org/lab-datasets/disambiguation/rich-author-disambiguation-data.zip

[4] http://dx.doi.org/10.7802/1234

[5] https://www.nlm.nih.gov/bsd/medline.html

[6] https://orcid.org/

[7] https://github.com/stanfordnlp/CoreNLP/blob/master/data/edu/stanford/nlp/patterns/surface/stopwords.txt

[8] https://tartarus.org/martin/PorterStemmer/

Ethnic Name Group (ENG) Tagging

This study assigns an ENG tag to a name instance in each labeled dataset using the author name ethnicity classification database, *Ethnea*, developed by Torvik and Agarwal (2016)[9]. *Ethnea* is a collection of more than 9 million author name instances that are tagged one of 26 ENG classes based on the name's association with national-level geo-locations[10]. For example, "Wang, Wei" is classified as 'Chinese' as it is most frequently associated with organizations in China. However, *Ethnea* makes no distinctions based on any anthropological, cultural, or linguistic characteristics of authors. Instead it relies entirely on observations of names and geo-locations of their frequently associated institutions so an author named 'Wang, Wei' who was born in the U.S. and has never visited China would still be assigned a 'Chinese name' tag. We link all four labelled datasets to *Ethnea* and, if a matched name is found, assign that name's ENG tag to all its observed instances. If a queried name does not have a match in *Ethnea*, we search again using only the surname and assign the modal *Ethnea* ENG tag associated with it to all its instances. Table 1 summarizes the frequencies and ratios of ENG tags assigned by *Ethnea* to author name instances in each labeled dataset.

Table 1 shows the list of ENGs in each labeled data. Small-sized ENGs are excluded from analysis because most name instances in those ENGs tend to belong to a single author while a few instances referring to other author(s). When randomly split into training and test subsets for machine learning, these instances do not produce negative pairs at all.

[Table 1]

Chinese names represent the majority of ENGs in three labeled data. This is because these labeled data were created from computer science papers where Chinese researchers are particularly large contributors. In addition, as the three datasets were designed to collate challenging names to disambiguate, Chinese names that tend to be more ambiguous than other ENGs were over-sampled (Müller, Reitz, & Roy, 2017). In contrast, 6,000 instances in UM-IRIS are evenly distributed over six ENGs. For validation and reuse, these labeled data with ENG tags are publicly available[11]. Note that the original KISTI contains 41,673 name instances, whereas the ENG-tagged KISTI has 41,605 instances. Such discrepancy occurs because this paper uses the revised version of KISTI that corrects record errors and duplicates in the original data (J. Kim, 2018).

Machine Learning Process

Machine learning methods for AND can be divided into two groups: author assignment and author grouping (Ferreira et al., 2012). While the former aims to assign an author name instance to one of pre-disambiguated author name clusters, the latter aims to group all and only instances that belong to the same authors. This study takes the latter approach in evaluating the effect of ENG on AND. Specifically, author

[9] https://databank.illinois.edu/datasets/IDB-9087546

[10] 26 ethnicities include: African, Arab, Baltic, Caribbean, Chinese, Dutch, English, French, German, Greek, Hispanic, Hungarian, Indian, Indonesian, Israeli, Italian, Japanese, Korean, Mongolian, Nordic, Polynesian, Romanian, Slav, Thai, Turkish, and Vietnamese. In *Ethnea*, some name instances are assigned two ethnicities (e.g., "Jane Kim" → Korean-English) if the surname and forename of an author name are associated frequently with different ethnicities.

[11] Download link TBA

name instances in each labeled dataset are pairwise compared to assess whether a given instance pair of plausibly represents the same author (a match) or not (a nonmatch). Although some scholars take a further step to cluster pairwise comparisons (e.g., J. Kim & Kim, 2018; Levin, Krawczyk, Bethard, & Jurafsky, 2012; Louppe et al., 2016; Alan Filipe Santana et al., 2017), this study only evaluates disambiguation performance at a pair level (i.e., classification), following the practice of previous AND studies (e.g., Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004; Song et al., 2015; Treeratpituk & Giles, 2009; Vishnyakova, Rodriguez-Esteban, Ozol, & Rinaldi, 2016).

As the first machine learning step, author name instances in each labeled dataset are randomly divided into training (50%) and test (50%) subsets. Then, instances in each subset are put into blocks in which all member instances share the same full surname and first forename initial (e.g., 'Wang, W'). Only instances in the same block are compared for disambiguation. This blocking is typical in AND studies because it reduces computational complexity with only slight performance degradation (K. Kim, Sefid, & Giles, 2017; Torvik & Smalheiser, 2009). Next, instance pairs in the same block are compared to establish their similarity over four other data features: author name, coauthor name(s), paper title, and publication venue. To quantify how much a pair is similar over a feature, this study calculates the cosine similarity of Term (*n*-gram) Frequency for each feature (Han, Zha, & Giles, 2005; J. Kim & Kim, In print; Levin et al., 2012; Louppe et al., 2016; A. F. Santana, Gonçalves, Laender, & Ferreira, 2015; Treeratpituk & Giles, 2009). Specifically, the string of a feature is converted into an array of 2~4-grams (e.g., author name 'Wang, Wei' → 'wa|an|ng|gw|we|ei|wan|ang|ngw|gwe|wei|wang|angw|ngwe|gwei'). After the conversion, two *n*-gram arrays of an instance pair are compared to produce a cosine similarity score for the feature.

Besides the four basic features, ENGs are used as a feature set for ENG-aware disambiguation. For this, especially, an instance pair's ENG is encoded into a binary value (i.e., one-hot encoding) for a pre-defined set of ENGs[12]. For example, in AMINER, a pair of name instances ('Wang, Wei' and 'Wang, W.') is assigned either 'Yes' or 'No' for each of five ethnicities – Chinese ('Yes'), English ('No'), Indian ('No'), German ('No'), and Hispanic ('No') – as shown in Table 1. Table 2 shows examples of the cosine similarity scores calculated over four features and ENG encoding results for instance pairs.

[Table 2]

We focus on four algorithms – Gradient Boosting, Logistic Regression, Naïve Bayes, and Random Forest – for supervised machine learning that have been widely used as baselines or best performing methods in AND studies (e.g., Han et al., 2004; J. Kim & Kim, In print; K. Kim, Sefid, Weinberg, & Giles, 2018; Louppe et al., 2016; Song et al., 2015; Torvik & Smalheiser, 2009; Treeratpituk & Giles, 2009; Vishnyakova et al., 2016; J. Wang et al., 2012). In the first scenario, they are trained on the list of similarity scores and labels, as shown in Table 2 to learn relative weights for features and an absolute weight or threshold for instance pairs to be disambiguated without considering ENGs (→ ENG-ignorant learning). In the second scenario, the same algorithms are trained on the list of similarity scores, ENGs, and labels (→ ENG-aware disambiguation). Here, the ethnic name partition adds more features (dimensions) to each instance pair's feature set, allowing algorithms to combine the similarities of the expanded features. The machine learning procedure is implemented using the python Scikit-learn package. For Gradient Boosting, 500 estimators are used with max depth=9 and learning rate = 0.125. For

[12] As only instances in the same block in which they share at least the same full name and first forename initial are compared, all the pairs in the block have the same ethnicity tag.

Logistic Regression, L2 Regularization with class weight = 1 is used. Gaussian Naïve Bayes with maximum likelihood estimator is used for Naïve Bayes. For Random Forest, 500 trees are used after a grid search.

Trained algorithmic models are applied to the instance pairs in test subsets in which the cosine similarity is calculated for the four basic features and, in the second scenario, ethnicities are encoded in the same fashion but explicitly include ENG information. As in Table 2, an algorithmic model receives a set of feature similarity scores and, if ENG-aware disambiguation is conducted, a list of encoded ENGs for an instance pair to output a binary classification decision (match or nonmatch). Once trained, each algorithm produces a single score that predicts the probability of an instance pair being negative (nonmatch). If the predicted probability is above a certain threshold (> 0.5), the pair is decided to be a nonmatch, whereas if below the threshold, a match.

Performance Evaluation

We evaluate each algorithm's classification results on reserved test subsets of each labeled dataset by calculating precision and recall for positive (P; match) and negative (N; nonmatch) pairs respectively. In addition, we calculate the F1 score as a harmonic mean of precision and recall.

Specifically, precision for positive pairs (Prec-Pos) measures how many predicted match pairs are correct ones (true positives; TP) over the total number of predicted match pairs that may contain correct match pairs (true positives; TP) and incorrect match pairs (false positives; FP). In contrast, recall for positive pairs (Rec-Pos) measures the ratio of correct match pairs (true positives; TP) over the total number of true match pairs that may be predicted correctly as match pairs (true positives; TP) or incorrectly as nonmatch pairs (false negatives; FN).

$$Precision\ Positive\ (PrecPos) = \frac{Number\ of\ Correctly\ Predicted\ Match}{Number\ of\ Predicted\ Match} = \frac{TP}{(TP + FP)} \quad (1)$$

$$Recall\ Positive\ (RecPos) = \frac{Number\ of\ Correctly\ Predicted\ Match}{Number\ of\ True\ Match} = \frac{TP}{(TP + FN)} \quad (2)$$

$$F1\ Positive = \frac{2 \times PrecPos \times RecPos}{PrecPos\ +\ RecPos} \quad (3)$$

Likewise, precision for negative pairs (Prec-Neg) measures how many predicted nonmatch pairs are correct ones (true negatives; TN) over the total number of predicted nonmatch pairs that may contain correct nonmatch pairs (true negatives; TN) and incorrect nonmatch pairs (false negatives; FN). In contrast, recall for negative pairs (Rec-Neg) measures the ratio of correct nonmatch pairs (true negatives; TN) over the total number of true nonmatch pairs that may be predicted correctly as nonmatch pairs (true negatives; TN) or incorrectly as match pairs (false positives; FP).

$$Precision\ Negative\ (PrecNeg) = \frac{Number\ of\ Correctly\ Predicted\ Nonmatch}{Number\ of\ Predicted\ Nonmatch} = \frac{TN}{(TN + FN)} \quad (4)$$

$$Recall\ Negative\ (RecNeg) = \frac{Number\ of\ Correctly\ Predicted\ Nonmatch}{Number\ of\ True\ Nonmatch} = \frac{TN}{(TN + FP)} \quad (5)$$

$$F1\ Negative = \frac{2 \times PrecNeg \times RecNeg}{PrecNeg\ +\ RecNeg} \quad (6)$$

The metrics are calculated on the entire set of test results for each labeled dataset. We separately calculate performance measures for difference ENGs rather than averaging them across multiple ethnicity groups.

Results

Cross-Data Performance Evaluation

Figure 1 shows disambiguation results on KISTI, reporting precision and recall *before* and *after* ENG-aware disambiguation by four algorithms – Gradient Boosting (GB), Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF). Figure 1a shows that when ENGs are included as features, the algorithms tend to produce better precision in the prediction of positive (match) pairs than when they are not considered. This is shown by black bars ('After') being higher than stripped bars ('Before') in Figure 1a. This observation indicates that ethnic name partitioning helps algorithms increase the ratio of TP among predicted positive pairs (= TP + FP). This can be confirmed by checking the numbers of true and false positive pairs in Table 3. For example, when trained only on the four basic (non-ENG) features, LR predicts that 76,201 (= TP + FP = 55,998 + 20,203) pairs refer to the same authors (match) and 73.49% of the predictions are right (= TP/(TP + FP)). After trained on the same but ENG-tagged data, however, it predicts 170,432 pairs to be match sets, increasing its prediction accuracy this time to 77.08%.

[Figure 1]

[Table 3]

Ethnic name partitioning also reduces the number of falsely predicted nonmatch cases (FN), increasing recall in Figure 1b. Performance gains by ENG-aware disambiguation are more pronounced for recall than for precision, as evidenced by larger differences between 'Before' and 'After' bars for recall (Figure 1b) than those for precision (Figure 1a). In other words, ENG aware disambiguation across four common algorithms appears to reduce false negative predictions more than true positive predictions, potentially providing better performance for applications (such as network analysis) that are particularly sensitive to biases due to erroneous "lumping" of name instances that actually refer to different individuals. The improvements in precision and recall together increase the F1 scores by ENG-aware disambiguation (Figure 1c).

ENG-aware disambiguation also does a better job of accurately predicting non-match (negative pair) cases. The 'After' bars are taller than those of 'Before' in Figure 1d. Unlike the positive pair prediction in which ENG-aware disambiguation works in favor of both precision and recall by all algorithms, however, the performance gains in precision for negative pairs come with slightly decreased recall by GB and LR in Figure 1e. This means that while disambiguation models by GB and LR trained on ENG-added features are good at increasing the numbers of true nonmatch pairs among predicted nonmatch pairs (= TN +FN), they incorrectly predict that true nonmatch pairs match (FP predictions) more frequently than when they are trained on the four basic features alone. Reduced recall for negative pair prediction is, however, offset by increased precision, leading to the F1 scores by ENG-aware disambiguation being better than those by ENG-blind one in Figure 1f. Meanwhile, NB and RF still obtain improvements in both precision and recall as well as F1.

Algorithmic performances are also enhanced by ENG-aware disambiguation on AMINER, GESIS, and UM-IRIS. Figure 2 ~ 4 report that the algorithms trained on ENG-tagged data perform better than those trained only on the basic features across almost all metrics for both positive and negative pairs. NB models prove the exception, producing worse results in recall for positive pairs and in precision for negative pairs after ethnic name partition. However, this degraded performance is offset by increased precision for positive pairs and increased recall for negative pairs, respectively, so the overall performance metric (F1), which equally weights precision and recall, indicates an overall improvement due to the inclusion of ENG features.

[Figure 2]

[Figure 3]

[Figure 4]

Performance Evaluation per ENG

ENG-aware disambiguation produces substantial improvements in both precision and recall for predicting match and nonmatch instance pairs in different labeled datasets. But are those improvements uniform across different ENGs? If not, a more nuanced approach to model evaluation may be necessary. To answer this question, we compare performance changes due to ENG-aware disambiguation within ENG groups. For this, precision, recall, and F1 scores for positive and negative pairs predicted by four algorithms are calculated separately for instance pairs that belong to the same ENG in each of four labeled data: 4 algorithms × 4 data = 16 evaluations. Presenting all the results at the same time would consume too much space in this paper. So, we present random forest (RF) predictions on the GESIS dataset as an illustration for the purposes of this discussion. Reports of other algorithms and data are presented in a supplementary document attached to this paper.

Figure 5 shows the by ENG performance metrics for the RF algorithm trained on GESIS with and without ENG-aware disambiguation. The ENG-aware disambiguation leads to better precision (positive pairs; Figure 5A) and recall (negative pairs; Figure 5E) for Chinese names but worse precision (positive pairs) and recall (negative pairs) for other ethnicities. In contrast, name disambiguation for Chinese names results in lower recall (positive pairs; Figure 4B) and precision (negative pairs; Figure 4D) than those for other ENGs. Similar patterns are observed for other algorithms tested on GESIS (see Figure S5~S8 in Supplementary Material). This suggests that the effect of ENG-aware disambiguation occurs in different ways for different ENGs. Thus, its application can be beneficial in some instances but detrimental in others. Variations in the effects of ENG-aware disambiguation on precision and recall for positive and negative pair prediction across ethnicity groups suggest that care must be taken to design disambiguation strategies that fit particular analytic or empirical needs.

[Figure 5]

These observations can be explained as follows. ENGs have different distributions of similarity scores over the four basic (non-ethnicity) features we use. Figure 6 presents the feature similarity score distributions per ENG for positive (left) and negative (right) pairs in the GESIS test data. Training and test subsets show similar distributions in each labeled dataset. For visual simplicity, a score is rounded up into nearest bins with intervals of 0.1 on *x*-axis and the ratios of the numbers of scores in the same bin over all scores are plotted on *y*-axis. A solid red line represents the distribution of all instance pairs regardless of ENG.

In Figure 6, each ENG has different distributions of, for example, 'COAUTHOR' similarity scores for both positive and negative pairs (Figure 6C and 6D). So, the four algorithms come to use different 'coauthor' similarity score distributions in ENG-aware disambiguation. Such heterogeneous distributions also occur for other features but with different variations of differences. For example, 'VENUE' distributions in Figure 6G and 6H differ less across ENGs than do 'COAUTHOR' distributions. Because ENG-aware disambiguation allows training and testing on different feature similarity score distributions for each ENG, the algorithms combine features using different weightings for each ethnicity, producing different predictions for name pairs with the same feature similarity scores but different ENG tags. In other words, this method takes into account the likelihood that researchers in different ENGs organize their scientific work differently, favoring distinct co-authorship and publication venue patterns. This also occurs in disambiguation of other labeled data, whose feature similarity score distributions are reported in Figure S17 ~ S20 in Supplementary Material.

[Figure 6]

Figure 5 also shows that some ENGs manifest substantial improvements in recall for positive pairs (Figure 5B) but degraded recall for negative pairs (Figure 5E). This might be explained in two ways. In our 'before' (ENG-unaware) case, algorithms combine features to produce per-feature weights for positive pairs based on feature similarity scores aggregated across multiple ENGs that can have very different feature distributions. Such aggregated distributions cannot effectively capture the single match patterns specific to each ENG, which seem to lead models to falsely predict positive pairs as negative ones (FN), reducing the recall for positive pairs. Conversely, increased recall for positive pairs after ENG-aware disambiguation means that the algorithms trained and tested on ENG-tagged data successfully produce per-feature weights optimized to each ENG, thus making better predictions that push up the recall scores for many ENGs.

Second, decreased negative pair recall after ENG-aware disambiguation means that the algorithms trained and tested on ENG-tagged data fail to produce proper per-feature weights for accurately predicting nonmatch for known negative pairs. When the algorithms are trained only on the four basic (non-ethnicity) features, they do a better job of predicting nonmatch pairs based on aggregated feature similarity distributions that are invariant across particular ENGs. In other words, feature distributions aggregated across ENGs appear to be more effective for predicting negative case pairs while ENG-aware disambiguation techniques more accurately capture positive pairs.

These observations imply that disambiguation models for positive pair prediction would be improved by ENG-aware procedures, while nonmatch patterns for negative pair prediction can aggregate across ENGs (J. Kim & Kim, 2018). Table 4 shows that in the GESIS training data, each ENG has different sizes of positive and negative (pairwise) pairs. CHINESE name instances produce the largest numbers of positive

(≈ 146K) and negative pairs (≈ 551K), while ITALIAN name instances generate around a few thousand positive and a few hundred negative pairs. In other training data, CHINESE pairs constitute substantially large proportions (KISTI: 71.28 % and AMINER: 91.26 %) or over one-third (UM-IRIS: 37.54 %) of all negative pairs, while other ENG pairs make up small or less-than-expected (approximately 17% per ENG in UM-IRIS) proportions. In contrast, the numbers of positive pairs are less concentrated (GESIS, KISTI, and AMINER) or more evenly distributed (UM-IRIS) for positive pairs than those for negative pairs.

[Table 4]

[Table 5]

[Table 6]

[Table 7]

As noted above for Figure 5 and observed in other labeled data (see Figure S1 ~ S16 in Supplementary Material), the algorithms work better in finding more true negative pairs even for non-CHINESE name pairs when they are trained on data in which ethnic name partitioning is not performed ('BEFORE') and, thus, negative pairs are dominated by CHINESE ones as shown in Table 4~6. This implies that the nonmatch patterns in CHINESE name pairs are applicable to predicting nonmatch pairs for other ENGs. In contrast, during ENG-aware disambiguation, the algorithms come to rely on the small-size negative pairs that may skew or distort true nonmatch patterns for some ENGs. This seems to result in the decreased recall in predicting negative pairs (i.e., many true negatives classified as false positives, which reduces precision for positive pair prediction), while increasing slightly precision in predicting negative pairs.

Despite the aforementioned conflicting changes in precision and recall per ENG, the overall performance by the four algorithms on the whole test set are shown in Figure 1~4 to substantially increase across the four labeled data after ethnic name partitioning is included in machine learning. One reason would be that performance gains outweigh losses at each ENG level overall. Another reason would be that especially for KISTI, AMINER, and GESIS, the improved performances in disambiguating CHINESE that constitute the majority of name instances may affect the overall evaluation results. As shown by the case of UM-IRIS in which ENG sizes are controlled to be equal, however, the overall performance improvements can be observed for all the ENGs by ENG-aware disambiguation. As such, this study illustrates that the ethnic name partition can be truly effective in improving disambiguation performances.

Discussion

These results suggest that AND tasks may produce better results by using ethnic name partition in machine learning. Considering that adding more features can improve generally machine learning performances, the enhanced disambiguation performances by ENG partitioning might not be a surprise. With that said, the real contribution of this study would be that it demonstrates many machine learning based disambiguation models have a potential to be improved by introducing ethnic name grouping into ambiguous data without additional collection of feature information.

To fully realize this potential, however, a few issues need to be addressed. First, ENG tagging can be a non-trivial task that requires an intricate algorithmic technique itself. Thanks to the ENG classification system developed and publicly shared by Torvik and Agarwal (2016), this study could assign ENGs to the names in four labeled data. Although *Ethnea* was constructed based on more than 9 million author name instances in PubMed, the world largest biomedicine library, it is unknown how well it can help us tag ENGs to names in other fields. Ideally, *Ethnea* may be updated regularly to reflect new author names entering bibliographic data in various fields. Practically, further research may be focused on finding out a set of ENGs that are most influential in improving disambiguation results and thus simplifying ENG tagging for author name disambiguation (e.g., CHINESE vs Non-CHINESE).

Second, the findings of this study were based on three labeled data (KISTI, AMINER, and GESIS) in which CHINESE names are dominant and the overall performance improvements were heavily affected by those for CHINESE name instances. To overcome such an imbalance of instance distribution in labeled data, a new labeled data (UM-IRIS) were created in a way that six ENGs have the same amount of ambiguous name instances. Disambiguation results from the new labeled data were in line with those from other three labeled data. In addition, all ENGs including CHINESE were able to obtain gains in disambiguation performances. But all these findings were obtained from small-sized labeled data, whether they are biased or controlled for ENG sizes. So, it is still unknown whether such improvements are achievable in author name disambiguation for large-scale bibliographic data in which ENG composition may be quite different from those in the labeled data used in this paper.

Another issue would be that there can be other features than the four used in this study that can lead ethnic name partition to different AND performances. For example, English authors may appear in publication records that are more complete in affiliation information and use more diverse title terms. Meanwhile, Chinese authors may tend to work with coauthors who have similar names in same institutions. Various features need to be explored to study further the impact of ethnic name partition on AND.

Fourth, ENG-aware disambiguation may be beneficial for positive pair prediction but not so much for negative pair prediction. This was illustrated in Figure 5 above and Figure S1 ~ S16 in Supplementary Material by the dramatically decreased recall in negative pair predictions for many ENGs. It was contrasted with the substantial increase of precision in positive pair prediction for those ENGs. This study speculates that by ethnic name partitioning, classifiers become stricter for CHINESE pairs while relaxed for other ENG pairs. In other words, a pairs of CHINESE instances that would be classified as 'match' before partitioning are classified as 'nonmatch' after partitioning (PREC-POS↑; REC-POS↓), while 'nonmatch' pairs of other ENG instances as matched ones (PREC-POS↓; REC-POS↑). This might be because while some CHINESE pairs sharing coauthor names, venue names, or title words refer to different authors, other ethnic names sharing the features are more likely to represent the same authors (see Figure 6 B, D, F, and H in which Chinese name pair share is denoted in square). During training, such different similarity patterns are mixed up before partitioning but distinguished after it. Another conjecture is that due to the relaxed classification after partitioning, true negative pairs of other-than-CHINESE ENGs are falsely classified as false positive pairs (PREC-POS↓; REC-NEG↓). As the sizes of negative pairs in most non-CHINESE ENGs are smaller than positive pairs (see Table 4~7), misclassified negative pairs have larger impacts on REC-NEG than on PREC-POS across the ENGs. But these conjectures are based on the observations on labeled data in which Chinese name instances are prevalent. Using an additional labeled data with controlled ENG sizes, however, the conjecture has been confirmed. But only six ENGs in a small dataset were considered for analysis. More ENGs need to be investigated to check if this conjecture holds good under the different combinations of ENGs.

Conclusion and Discussion

This study evaluated the effects ethnic name partitioning has on author name disambiguation (AND) using machine learning methods. For this, author name instances in four labeled datasets were disambiguated under two scenarios. First, similarity scores of instance pairs over four basic features – author name, coauthor names, paper title, and publication venue – were used to train and test disambiguation algorithms. Second, in addition to the basic features, ethnic name group (ENGs) were tagged to name instances to allow algorithms to build models that are optimized to each ENG. Comparisons of disambiguation performances before and after ENG-aware disambiguation showed that using ethnic name partition can substantially improve algorithmic performances. Such performance improvements occurred across all ENGs, although performance gains and losses at each ENG level were observed in different ways depending on the types of measures – precision or recall – and target classifications – positive (match) or negative (nonmatch) pairs.

As detailed in the discussion above, ethnic name partition requires further research for us to understand better its impact on author name disambiguation and apply it to disambiguation tasks for digital libraries that are struggling with authority control over fast-growing ambiguous author names. This study is expected to motivate scholars and practitioners to study toward that direction by demonstrating the potential of ENG-aware disambiguation in improving disambiguation performances.

Acknowledgements

This paper was supported by grants from the National Science Foundation (#1917663, #1760609, and #1561687). The authors would like to thank anonymous reviewers for their comments to improve the paper.

References

Chin, A. W.-S., Juan, Y.-C., Zhuang, Y., Wu, F., Tung, H.-Y., Yu, T., . . . et al. (2013). *Effective string processing and matching for author disambiguation*. Paper presented at the Proceedings of the 2013 KDD Cup 2013 Workshop, Chicago, Illinois. https://doi.org/10.1145/2517288.2517295

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. F. (2010). An Unsupervised Heuristic-Based Hierarchical Method for Name Disambiguation in Bibliographic Citations. *Journal of the American Society for Information Science and Technology, 61*(9), 1853-1870. doi:10.1002/asi.21363

Deville, P., Wang, D. S., Sinatra, R., Song, C. M., Blondel, V. D., & Barabási, A.-L. (2014). Career on the Move: Geography, Stratification, and Scientific Impact. *Scientific Reports, 4*. doi:10.1038/srep04770

Fegley, B. D., & Torvik, V. I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLOS ONE, 8*(7). doi:10.1371/journal.pone.0070299

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. *Sigmod Record, 41*(2), 15-26.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. F. (2014). Self-Training Author Name Disambiguation for Information Scarce Scenarios. *Journal of the Association for Information Science and Technology, 65*(6), 1257-1278. doi:10.1002/asi.22992

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). *Two Supervised Learning Approaches for Name Disambiguation in Author Citations.* Paper presented at the Jcdl 2004: Proceedings of the Fourth Acm/Ieee Joint Conference on Digital Libraries, Tucson, Arizona.

Han, H., Zha, H. Y., & Giles, C. L. (2005). Name disambiguation in author citations using a K-way spectral clustering method. *Proceedings of the 5th Acm/Ieee Joint Conference on Digital Libraries, Proceedings*, 334-343. doi:Doi 10.1145/1065385.1065462

Harzing, A. W. (2015). Health warning: might contain multiple personalities-the problem of homonyms in Thomson Reuters Essential Science Indicators. *Scientometrics, 105*(3), 2259-2270. doi:10.1007/s11192-015-1699-y

Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review, 32*, e22. doi:10.1017/S0269888917000182

Hussain, I., & Asghar, S. (2018). DISC: Disambiguating homonyms using graph structural clustering. *Journal of Information Science, 44*(6), 830-847. doi:10.1177/0165551518761011

Islamaj Dogan, R., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database : the journal of biological databases and curation, 2009*, bap018-bap018. doi:10.1093/database/bap018

Kang, I. S., Kim, P., Lee, S., Jung, H., & You, B. J. (2011). Construction of a Large-Scale Test Set for Author Disambiguation. *Information Processing & Management, 47*(3), 452-465. doi:10.1016/j.ipm.2010.10.001

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics, 103*(3), 1061-1071. doi:10.1007/s11192-015-1580-z

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: A case of DBLP. *Scientometrics, 116*(3), 1867-1886. doi:10.1007/s11192-018-2824-5

Kim, J., & Diesner, J. (2015). The effect of data pre-processing on understanding the evolution of collaboration networks. *Journal of Informetrics, 9*(1), 226-236. doi:10.1016/j.joi.2015.01.002

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology, 67*(6), 1446-1461. doi:10.1002/asi.23489

Kim, J., & Owen-Smith, J. (In print). ORCID-linked labeled data for evaluating author name disambiguation at scale. Scientometrics. doi:10.1007/s11192-020-03826-6

Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics, 117*(1), 511-526. doi:10.1007/s11192-018-2865-9

Kim, J., & Kim, J. (2020). Effect of Forename String on Author Name Disambiguation. *Journal of the Association for Information Science and Technology, 71*(7), 839-855. doi:10.1002/asi.24298

Kim, K., Sefid, A., & Giles, C. L. (2017). Scaling Author Name Disambiguation with CNF Blocking. *arXiv preprint arXiv:1709.09657*.

Kim, K., Sefid, A., Weinberg, B. A., & Giles, C. L. (2018). *A Web Service for Author Name Disambiguation in Scholarly Databases.* Paper presented at the 2018 IEEE International Conference on Web Services (ICWS).

Lerchenmueller, M. J., & Sorenson, O. (2016). Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists. *PLOS ONE, 11*(7), e0158731. doi:10.1371/journal.pone.0158731

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-Based Bootstrapping for Large-Scale Author Disambiguation. *Journal of the American Society for Information Science and Technology, 63*(5), 1030-1047. doi:10.1002/asi.22621

Ley, M. (2009). DBLP: some lessons learned. *Proceedings of the VLDB Endowment, 2*(2), 1493-1500.

Louppe, G., Al-Natsheh, H. T., Susik, M., & Maguire, E. J. (2016). ENG Sensitive Author Disambiguation Using Semi-supervised Learning. *Knowledge Engineering and Semantic Web, Kesw 2016, 649*, 272-287. doi:10.1007/978-3-319-45880-9_21

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics, 7*(4), 767-773. doi:10.1016/j.joi.2013.06.006

Momeni, F., & Mayr, P. (2016). *Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names.* Paper presented at the 20th international Conference on Theory and Practice of Digital Libraries (TPDL 2016), Hannover, Germany.

Müller, M. C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics, 111*(3), 1467-1500. doi:10.1007/s11192-017-2363-5

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2015). On the Combination of Domain-Specific Heuristics for Author Name Disambiguation: The Nearest Cluster Method. *International Journal on Digital Libraries, 16*(3-4), 229-246. doi:10.1007/s00799-015-0158-y

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2017). Incremental author name disambiguation by exploiting domain-specific heuristics. *Journal of the Association for Information Science and Technology, 68*(4), 931-945. doi:10.1002/asi.23726

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2019). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*. doi:10.1177/0165551519888605

Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics, 107*(3), 1283-1298. doi:10.1007/s11192-016-1892-7

Shin, D., Kim, T., Choi, J., & Kim, J. (2014). Author Name Disambiguation Using a Graph Model with Node Splitting and Merging Based on Bibliographic Information. *Scientometrics, 100*(1), 15-50. doi:10.1007/s11192-014-1289-4

Smalheiser, N. R., & Torvik, V. I. (2009). Author Name Disambiguation. *Annual Review of Information Science and Technology, 43*, 287-313.

Song, M., Kim, E. H. J., & Kim, H. J. (2015). Exploring Author Name Disambiguation on PubMed-Scale. *Journal of Informetrics, 9*(4), 924-941. doi:10.1016/j.joi.2015.08.004

Strotmann, A., & Zhao, D. Z. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology, 63*(9), 1820-1833. doi:10.1002/asi.22695

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). *ArnetMiner: extraction and mining of academic social networks*. Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA.

Torvik, V. I., & Agarwal, S. (2016). *Ethnea: An instance-based ENG classifier based on geo-coded author names in a large-scale bibliographic database*. Paper presented at the Library of Congress International Symposium on Science of Science, Washington D.C. http://hdl.handle.net/2142/88927

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *Acm Transactions on Knowledge Discovery from Data, 3*(3). doi:10.1145/1552303.1552304

Treeratpituk, P., & Giles, C. L. (2009). Disambiguating Authors in Academic Publications using Random Forests. *JCDL 2009: Proceedings of the 2009 Acm/Ieee Joint Conference on Digital Libraries*, 39-48.

Vishnyakova, D., Rodriguez-Esteban, R., Ozol, K., & Rinaldi, F. (2016, dec). *Author Name Disambiguation in MEDLINE Based on Journal Descriptors and Semantic Types.* Paper presented at the Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), Osaka, Japan.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics, 93*(2), 391-411. doi:10.1007/s11192-012-0681-1

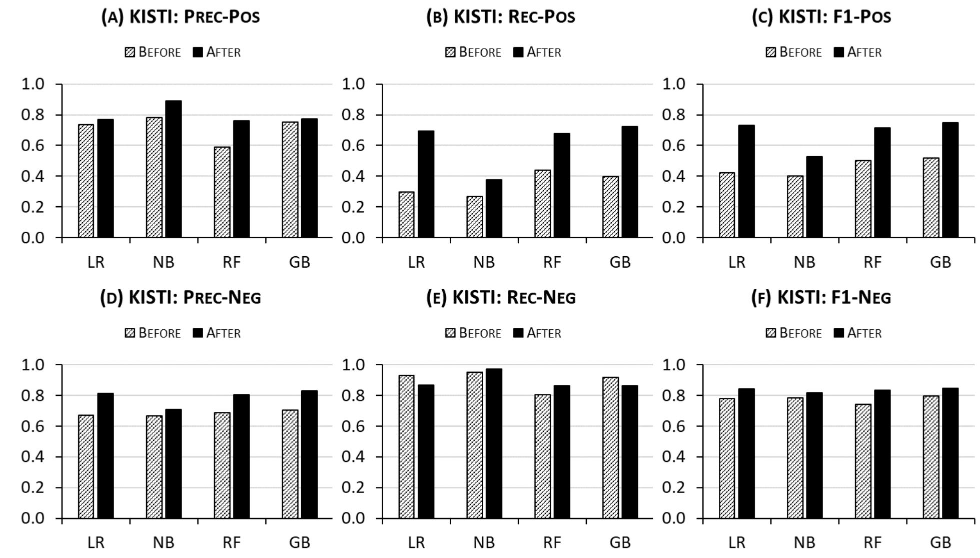
Wang, X., Tang, J., Cheng, H., & Yu, P. S. (2011). *ADANA: Active Name Disambiguation*. Paper presented at the 2011 IEEE 11th International Conference on Data Mining.

Wu, H., Li, B., Pei, Y. J., & He, J. (2014). Unsupervised author disambiguation using Dempster-Shafer theory. *Scientometrics, 101*(3), 1955-1972. doi:10.1007/s11192-014-1283-x

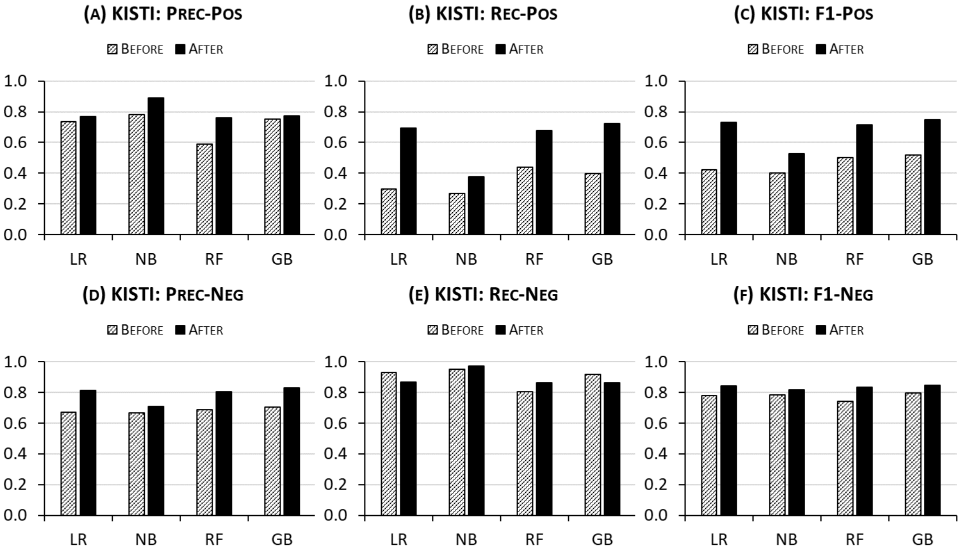
Wu, J., & Ding, X. H. (2013). Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics, 96*(3), 683-697. doi:10.1007/s11192-013-0978-8

Zhao, Z., Rollins, J., Bai, L., & Rosen, G. (2017, Oct. 2017). *Incremental Author Name Disambiguation for Scientific Citation Data.* Paper presented at the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Zhu, J., Wu, X., Lin, X., Huang, C., Fung, G. P. C., & Tang, Y. (2018). A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering. *Scientometrics, 114*(3), 781-794. doi:10.1007/s11192-017-2611-8



ASI_24459_figure1.tiff



(a) KISTI: Prec-Pos

(b) KISTI: Rec-Pos

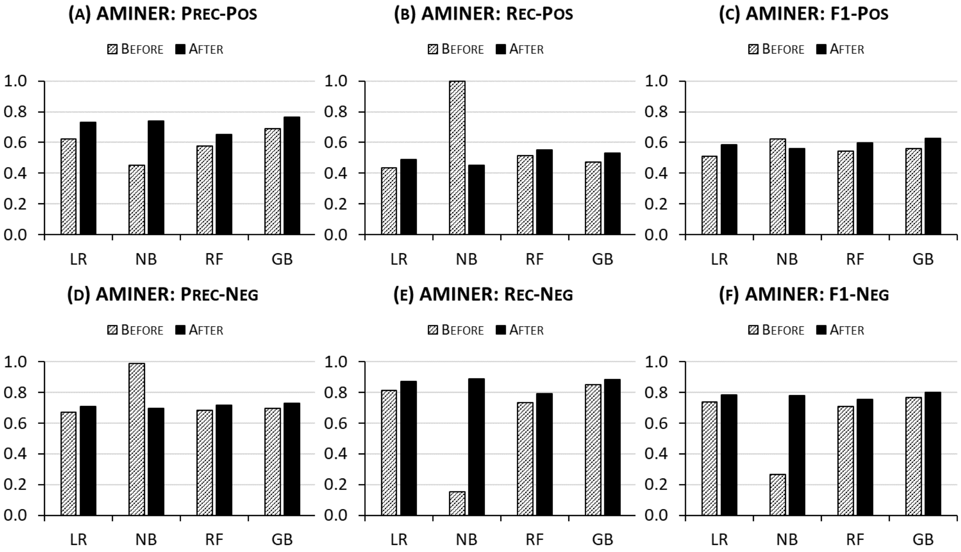
(c) KISTI: F1-Pos

(d) KISTI: Prec-Neg

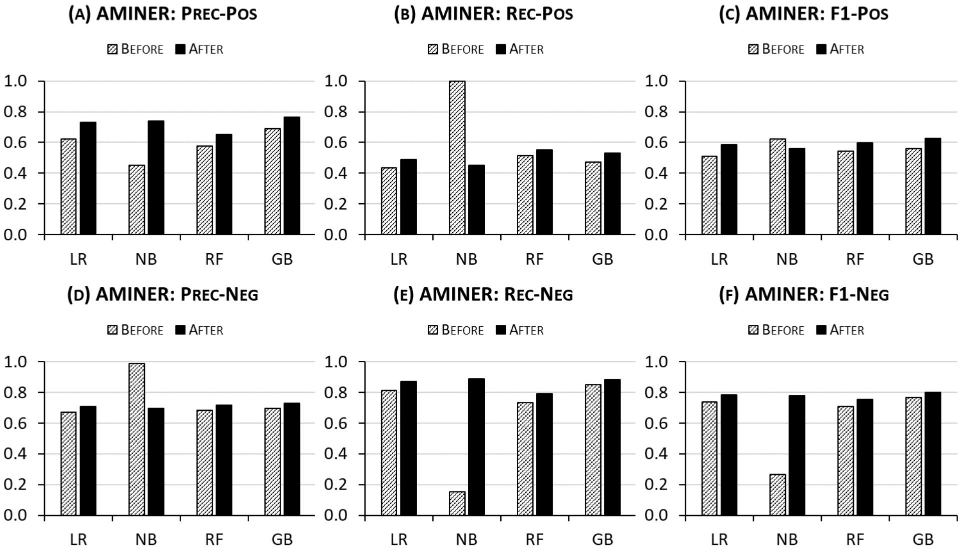
(e) KISTI: Rec-Neg

(f) KISTI: F1-Neg

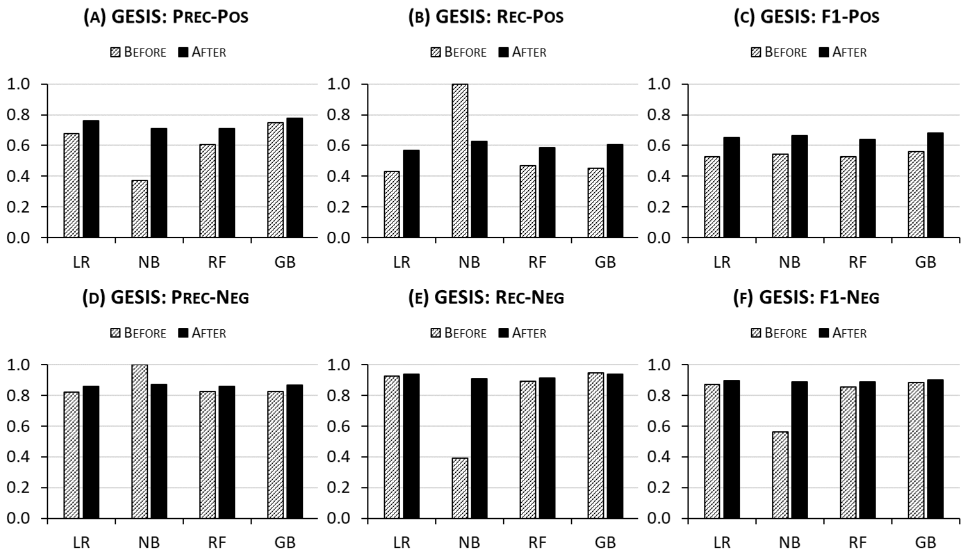
ASI_24459_figure1.tiff



ASI_24459_figure2.tiff



ASI_24459_figure2.tiff



(A) GESIS: Prec-Pos

(B) GESIS: Rec-Pos

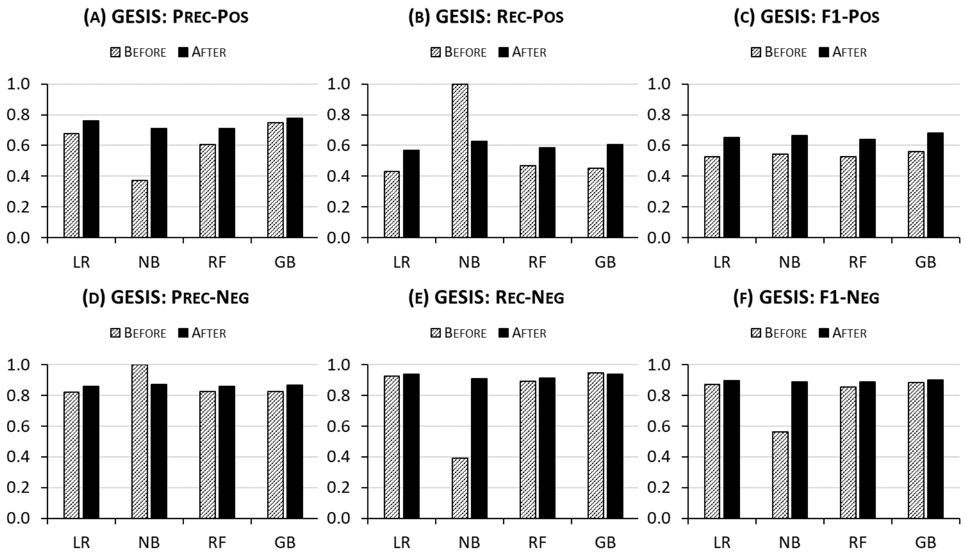
(C) GESIS: F1-Pos

(D) GESIS: Prec-Neg

(E) GESIS: Rec-Neg

(F) GESIS: F1-Neg

ASI_24459_figure3.tiff



(A) GESIS: PREC-POS

(B) GESIS: REC-POS

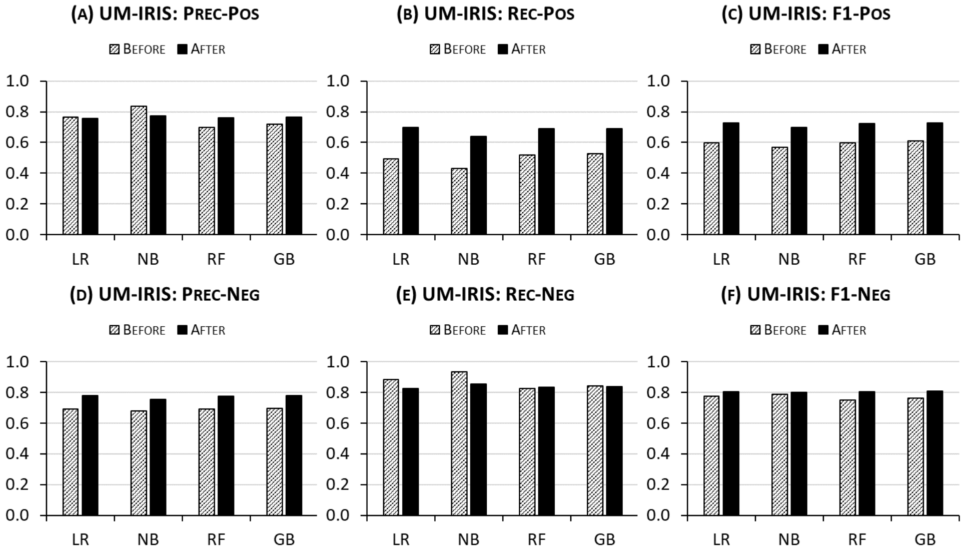
(C) GESIS: F1-POS

(D) GESIS: PREC-NEG

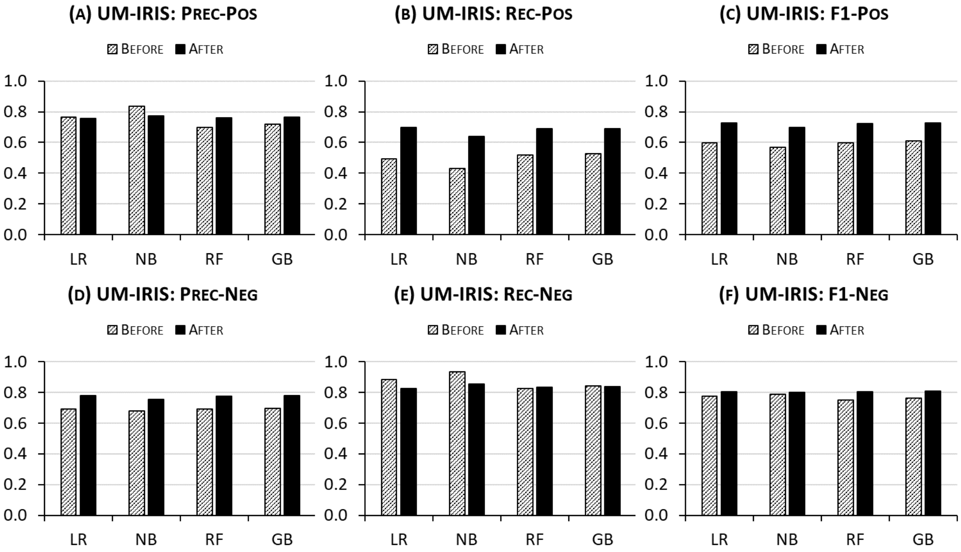
(E) GESIS: REC-NEG

(F) GESIS: F1-NEG

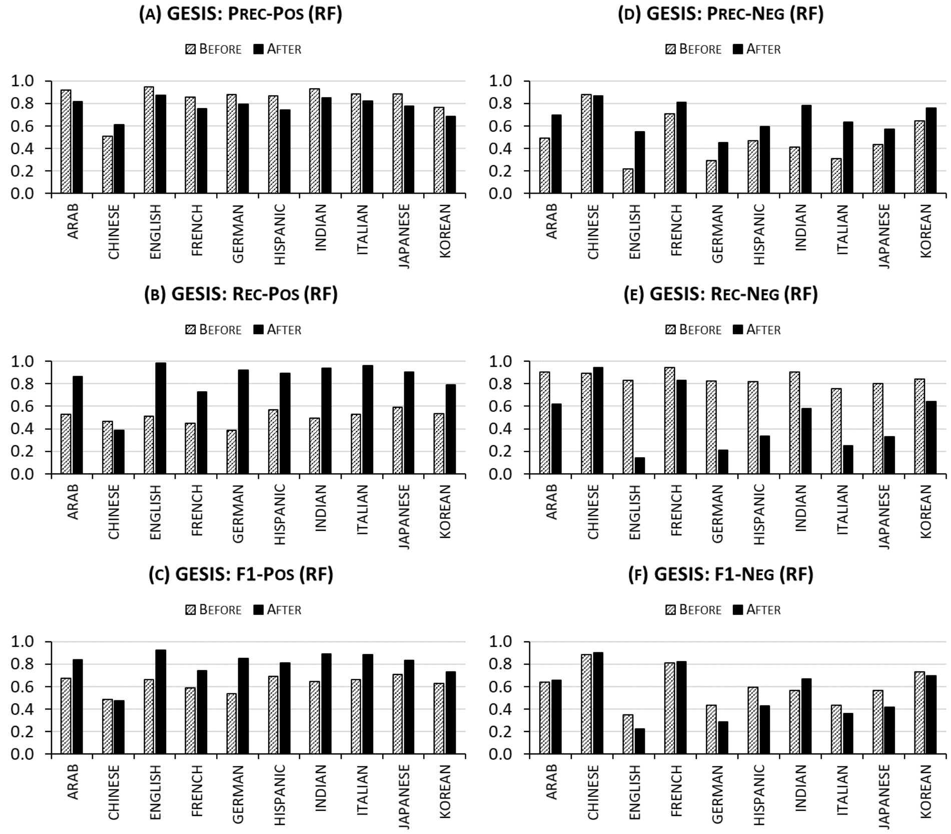
ASI_24459_figure3.tiff



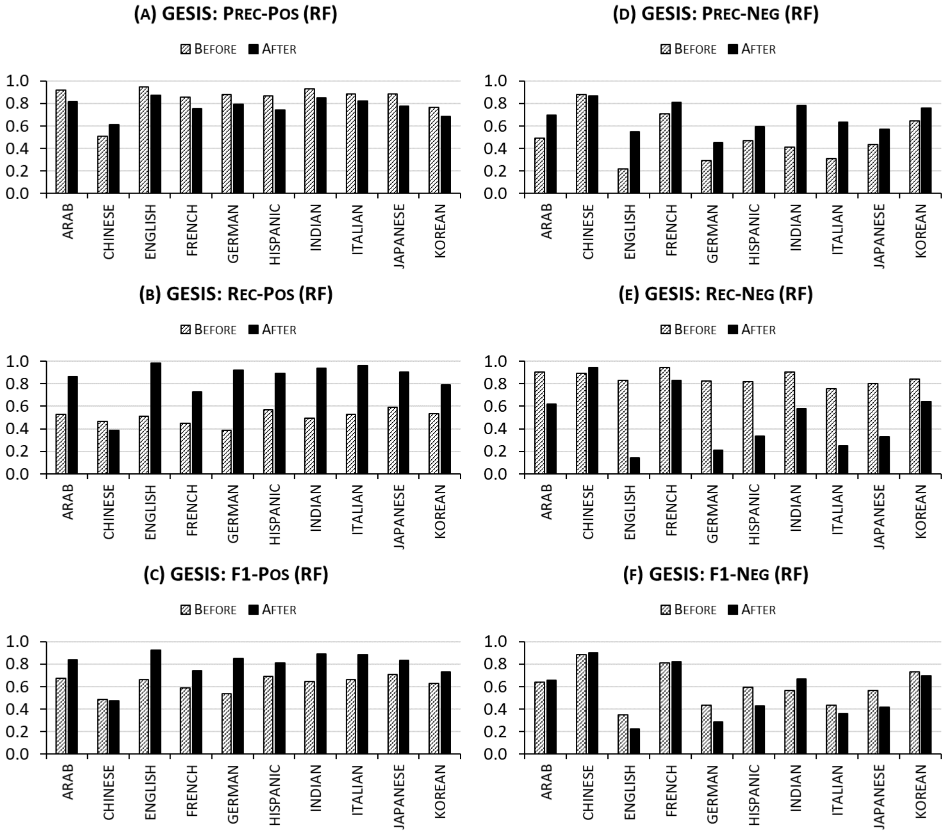
ASI_24459_figure4_rev.tif



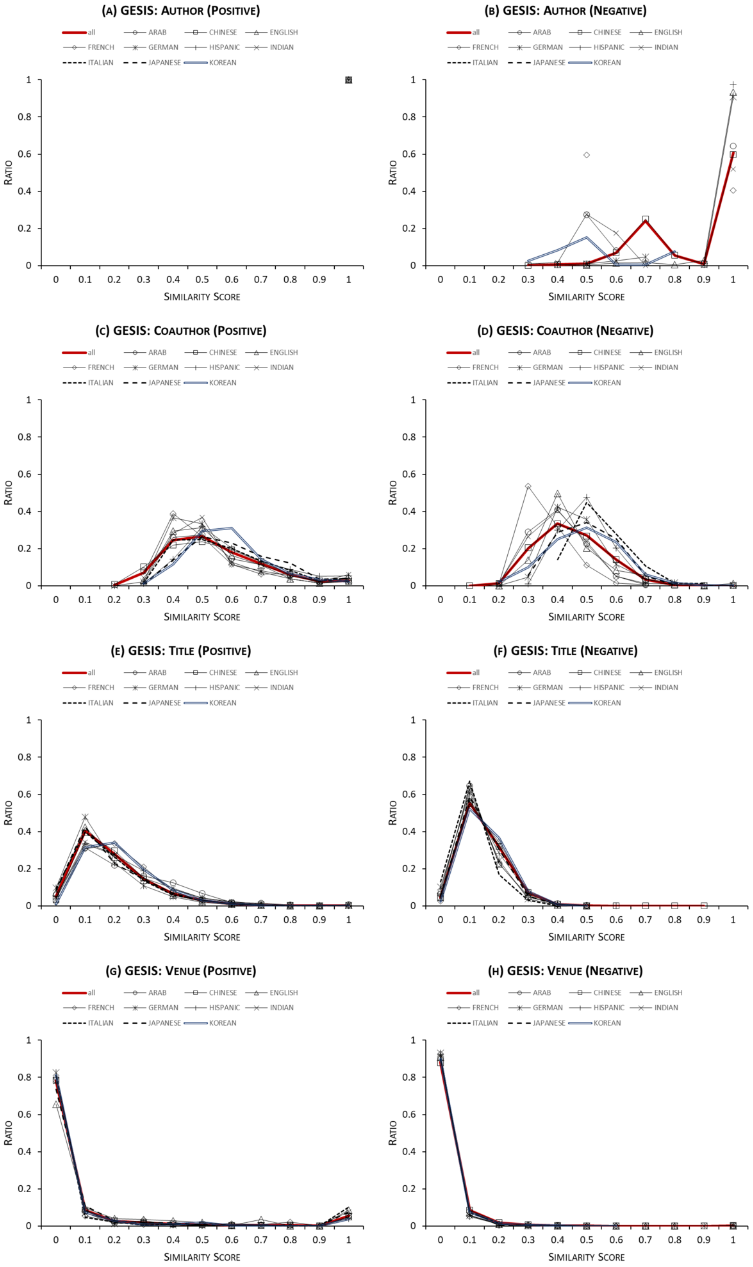
ASI_24459_figure4_rev.tif



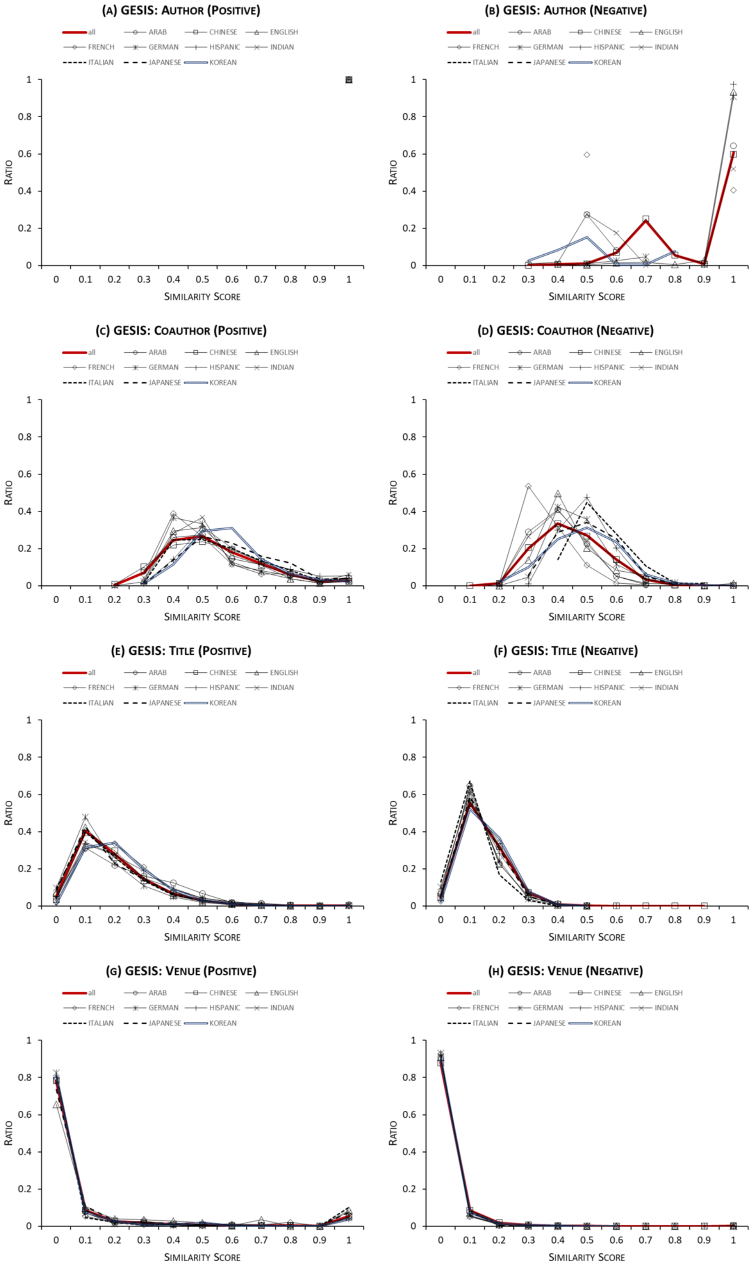
ASI_24459_figure5.tiff



ASI_24459_figure5.tiff



ASI_24459_figure6.tiff



ASI_24459_figure6.tiff

Table 1: Summary of Ethnic Name Group (ENG) Frequencies in Labeled Data

Table 2: A Mock-Up Example of Cosine Similarity Scores for Instance Pairs over Four Features and ENG Encoding

Figure 1: Disambiguation Performances ‘Before’ Versus ‘After’ ENG-Aware Disambiguation on KISTI

Table 3: Numbers of Correctly or Incorrectly Predicted Pairs for Positive and Negative Pairs by Four Algorithms on KISTI Test Data

Figure 2: Disambiguation Performances ‘Before’ Versus ‘After’ ENG-Aware Disambiguation on AMINER

Figure 3: Disambiguation Performances ‘Before’ Versus ‘After’ ENG-Aware Disambiguation on GESIS

Figure 4: Disambiguation Performances ‘Before’ Versus ‘After’ ENG-Aware Disambiguation on UM-IRIS

Figure 5: Disambiguation Performances per ENG ‘Before’ Versus ‘After’ ENG-Aware Disambiguation by Random Forest on GESIS

Figure 6: Feature Similarity Score Distributions per ENG for Positive and Negative Pairs in GESIS Test Data

Table 4: Distributions of Positive and Negative Name Instance Pairs per ENG in GESIS Training Data

Table 5: Distributions of Positive and Negative Name Instance Pairs per ENG in KISTI Training Data

Table 6: Distributions of Positive and Negative Name Instance Pairs per ENG in AMINER Training Data

Table 7: Distributions of Positive and Negative Name Instance Pairs per ENG in UM-IRIS Training Data

Table 1: Summary of Ethnic Name Group (ENG) Frequencies in Labeled Data

Table 2: A Mock-Up Example of Cosine Similarity Scores for Instance Pairs over Four Features and ENG Encoding

Figure 1: Disambiguation Performances ‘Before’ Versus ‘After’ ENG-Aware Disambiguation on KISTI

Table 3: Numbers of Correctly or Incorrectly Predicted Pairs for Positive and Negative Pairs by Four Algorithms on KISTI Test Data

Figure 2: Disambiguation Performances ‘Before’ Versus ‘After’ ENG-Aware Disambiguation on AMINER

Figure 3: Disambiguation Performances ‘Before’ Versus ‘After’ ENG-Aware Disambiguation on GESIS

Figure 4: Disambiguation Performances ‘Before’ Versus ‘After’ ENG-Aware Disambiguation on UM-IRIS

Figure 5: Disambiguation Performances per ENG ‘Before’ Versus ‘After’ ENG-Aware Disambiguation by Random Forest on GESIS

Figure 6: Feature Similarity Score Distributions per ENG for Positive and Negative Pairs in GESIS Test Data

Table 4: Distributions of Positive and Negative Name Instance Pairs per ENG in GESIS Training Data

Table 5: Distributions of Positive and Negative Name Instance Pairs per ENG in KISTI Training Data

Table 6: Distributions of Positive and Negative Name Instance Pairs per ENG in AMINER Training Data

Table 7: Distributions of Positive and Negative Name Instance Pairs per ENG in UM-IRIS Training Data

Title: Ethnicity-Based Name Partitioning for Author Name Disambiguation Using Supervised Machine Learning

Authors: Jinseok Kim, Jenna Kim, and Jason Owen-Smith

1. Jinseok Kim

Institute for Research on Innovation & Science, Survey Research Center, Institute for Social Research
University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104
jinseokk@umich.edu
ORCID id: 0000-0001-6481-2065

2. Jenna Kim

School of Information Sciences
University of Illinois at Urbana-Champaign
501 E. Daniel Street, Champaign, IL U.S.A. 61820
jkim682@illinois.edu

ORCID id: 0000-0001-7438-448X

3. Jason Owen-Smith

Department of Sociology, Institute for Social Research, University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104-2910
734-936-0463
jdos@umich.edu
ORCID id: 0000-0001-8007-1121

[Corresponding Author Information]

Jinseok Kim

Institute for Research on Innovation & Science, Survey Research Center, Institute for Social Research
University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104-2910
734-763-4994
jinseokk@umich.edu

[Financial Support Information]

This paper was supported by grants from the National Science Foundation (#1917663, #1760609, and #1561687).

Abstract

In several author name disambiguation studies, some ethnic name groups such as East Asian names are reported to be more difficult to disambiguate than others. This implies that disambiguation approaches might be improved if ethnic name groups are distinguished before disambiguation. We explore the potential of ethnic name partitioning by comparing performance of four machine learning algorithms trained and tested on the entire data or specifically on individual name groups. Results show that ethnicity-based name partitioning can substantially improve disambiguation performance because the individual models are better suited for their respective name group. The improvements occur across all ethnic name groups with different magnitudes. Performance gains in predicting matched name pairs outweigh losses in predicting nonmatched pairs. Feature (e.g., coauthor name) similarities of name pairs vary across ethnic name groups. Such differences may enable the development of ethnicity specific feature weights to improve prediction for specific ethic name categories. These findings are observed for three labeled data with a natural distribution of problem sizes as well as one in which all ethnic name groups are controlled for the same sizes of ambiguous names. This study is expected to motive scholars to group author names based on ethnicity prior to disambiguation.

Keywords: supervised machine learning; author name disambiguation; feature engineering, name ethnicity; data partition

Introduction and Background

A big challenge in managing digital libraries is that author names in bibliographic data are ambiguous because many authors have the same names (homonyms) or variant names are recorded for the same authors (synonyms). One study estimates that about two-thirds of author names in PubMed, the largest biomedicine digital library, are vulnerable to either or both of these two ambiguity types (Torvik & Smalheiser, 2009). Research findings obtained by mining bibliographic data can be distorted by merged and/or split author identities due to incorrect disambiguation (Fegley & Torvik, 2013; J. Kim & Diesner, 2015, 2016; Schulz, 2016). In addition, digital library users query author names most frequently (Islamaj Dogan, Murray, Névéol, & Lu, 2009). This means that the users will receive inaccurate information about research production, citation, and collaboration for authors if author name ambiguity is not properly resolved (Harzing, 2015; Strotmann & Zhao, 2012).

To address the challenge, researchers have proposed a variety of author name disambiguation (AND, hereafter) methods. Some scholars have used heuristics such as string-based matching (e.g., names that have the same full surname and forename initials are assumed to represent the same author), which is the most widely used approach in bibliometrics (Milojević, 2013). Others have developed rule-based programming and supervised/unsupervised machine learning techniques, as systemically reviewed in several papers (Ferreira, Gonçalves, & Laender, 2012; Hussain & Asghar, 2017; Sanyal, Bhowmick, & Das, 2019; Smalheiser & Torvik, 2009). In industry, several bibliographic data providers such as DBLP, Scopus, and Web of Science have disambiguated author names to improve their service quality (Kawashima & Tomizawa, 2015; J. Kim, 2018; Ley, 2009; Zhao, Rollins, Bai, & Rosen, 2017), while others still rely on the name string matching to output author-related search results.

Despite the differences in methods and datasets, a few AND studies have observed that some ethnic name groups (ENG, hereafter) (e.g., Chinese names) are more difficult to disambiguate than others (Deville et

al., 2014; J. Kim & Diesner, 2016; Strotmann & Zhao, 2012; Torvik & Smalheiser, 2009; J. Wu & Ding, 2013). This implies that author names may be better disambiguated if their associated ethnicities are considered as inputs in disambiguation models. But this possibility has been little explored. First, the observations made in several studies that certain ethnic names are harder to disambiguate are based on post-hoc evaluations of AND results. In other words, many of those studies did not integrate ethnic name partitions during machine learning. A very small number of studies have divided names into subgroups in their disambiguation model building (Chin et al., 2013; Louppe, Al-Natsheh, Susik, & Maguire, 2016) and evaluation process (Lerchenmueller & Sorenson, 2016). But their ethnic name categories are limited in number (e.g., dichotomy of Chinese vs non-Chinese; Caucasian, Asian, and Hispanic) or mixed up with racial distinctions based on the U.S. Social Security information (e.g., White, Black, Hispanic, Asian, etc.). Such racial classifications can be inappropriate for bibliographic data in which author names come from diverse regions around the world. In addition, those studies have typically used a single labeled data source, which makes it hard to expand and generalize their findings to other AND scenarios.

This study aims to empirically evaluate the effect ethnic name partitioning has on AND. In this study, author name disambiguation is a task to assign either 'match' or 'nonmatch' label to a pair of author name instances. For this, specifically, name instances are grouped into a block that share the same first forename initial and full surname and pairwisely compared with the block for their similarities over a set of features (e.g., coauthor name) to produce similarity scores. Machine learning algorithms combine the scores to learn weights of each feature to decide if a given pair of instances to refer to the same author or not. Although our work is motivated by the studies reviewed above and follows their common data pre-processing, blocking and machine learning steps, this paper differs from them in three important ways. First, this study evaluates AND performance by four different machine learning algorithms applied to four different labeled datasets before and after inclusion of a standard ethnic name group partition. Here, a name instance is assigned to an ethnic name group based on a name ethnicity classification system, *Ethnea*. Second, unlike traditional labeled data in which a specific ENG (i.e., Chinese) dominates, this study disambiguates new labeled data in which all ENGs are controlled to have the same numbers of instances, to demonstrate that performance changes induced by ethnic name partitioning may not be solely due to the well-known relationship between the number of cases and their ambiguity (more names, more ambiguity). Third, this study shows that different combinations of features (e.g., coauthor name and title words) appear to be related to AND performance for different ENGs suggesting future directions to further improve AND performance with ambiguous ethnic group names. The findings of this study can provide practical insights to researchers and practitioners who handle authority control in digital libraries. In the following sections, details on labeled data and setups for machine learning are described.

Method

Labeled Data and Pre-Processing

To measure the effect of ethnic name partition on machine learning for AND, this study disambiguates names in four labeled datasets – KISTI, AMINER, GESIS, and UM-IRIS. The first three datasets have been used in many AND studies to train and test machine learning algorithms (Cota, Ferreira, Nascimento, Gonçalves, & Laender, 2010; Ferreira, Veloso, Gonçalves, & Laender, 2014; Hussain & Asghar, 2018; J. Kim & Kim, 2018, In print; Momeni & Mayr, 2016; Alan Filipe Santana, Gonçalves, Laender, & Ferreira, 2017; Shin, Kim, Choi, & Kim, 2014; H. Wu, Li, Pei, & He, 2014; Zhu et al., 2018).

The last one is added to investigate how the ethnic name partition affects AND under the condition in which all ENGs are constrained to have the same numberss of ambigous name instances[1].

KISTI: Scientists at the Korea Institute of Science & Technology Information (KISTI) and Kyungsung University in Korea constructed this labeled dataset. It is made up of 41,673 author name instances that belong to 6,921 unique authors (Kang, Kim, Lee, Jung, & You, 2011)[2].

AMINER: Researchers in China and U.S. collaborated to create this labeled data to build and evaluate AND models for a computer science digital library, AMiner (Tang et al., 2008; X. Wang, Tang, Cheng, & Yu, 2011)[3]. It consists of 7,528 author name instances that refer to 1,546 unique authors.

GESIS: Scholars at the Leibniz Institute for the Social Sciences (GESIS) in Germany produced this labeled data. It contains author name instances of 5,408 unique authors (Momeni & Mayr, 2016) [4]. This study reuses the 'Evaluation Set' (29,965 author name instances of 2,580 unique authors) but with a few enhancements (J. Kim & Kim, 2020). Each author name instance is converted into the 'surname, forename' format and, through linking GESIS to its base DBLP data, is associated with the title of the paper in which it appears and the name of the conference or journal where the paper is published.

UM-IRIS: This dataset was generated by the researchers at the University of Michigan Institute for Research on Innovation & Science (UM-IRIS) and the University of Illinois through matching selected name instances in publication records to an authority database, ORCID (Kim & Owen-Smith, in print). First, author full names (e.g., 'Brown, Michael') that appear 50 times or more in MEDLINE-indexed publications published between 2000 and 2019 were listed[5]. Then, all instances of each selected full name (e.g., 158 instances of 'Brown, Michael' in MEDLINE) and their associated publication metadata were compared to 6 million researcher profiles in ORCID[6]. If an instance had a single match in the publication list of an ORCID researcher profile (matching on full name, paper title, and publication venue), the matched researcher's ORCID id was assigned as an author label to the instance. Next, among the ORCID id-linked instances, those whose full names are associated with 5 or more ORCID ids (e.g., 6 unique ORCID researchers share the name 'Brown, Michael' which appear 158 times in MEDLINE) were randomly selected to produce 1,000 name instances for each of six ENGs. The resulting data contain 6,000 instances of 822 authors.

Four features – author name, coauthor name(s), paper title, and publication venue - are used as machine learning features because they have been widely used in algorithmic AND studies (Schulz, 2016; Song, Kim, & Kim, 2015) and are commonly available in the four labeled datasets. The string of each feature is stripped of non-alphabetical characters, converted into ASCII format, and lowercased. For title words, common English words like 'the' and 'to' are removed (i.e., stop-word listed) using the dictionary in Stanford NLP[7] and stemmed (e.g., 'solution' → 'solut') using the Porter's algorithm[8]. Name instances in KISTI are converted into the full surname and first forename initial format ('Wang, Wei' → 'Wang, W') to make them more ambiguous (see J. Kim & Kim, 2020).

[1] This new labeled dataset was created following the idea of a reviewer who suggested that the impact of ethnic name partition on AND may be confounded by the size differences of ambiguous names.

[2] http://www.lbd.dcc.ufmg.br/lbd/collections/disambiguation/DBLP.tar.gz/at_download/file

[3] http://arnetminer.org/lab-datasets/disambiguation/rich-author-disambiguation-data.zip

[4] http://dx.doi.org/10.7802/1234

[5] https://www.nlm.nih.gov/bsd/medline.html

[6] https://orcid.org/

[7] https://github.com/stanfordnlp/CoreNLP/blob/master/data/edu/stanford/nlp/patterns/surface/stopwords.txt

[8] https://tartarus.org/martin/PorterStemmer/

Ethnic Name Group (ENG) Tagging

This study assigns an ENG tag to a name instance in each labeled dataset using the author name ethnicity classification database, *Ethnea*, developed by Torvik and Agarwal (2016)[9]. *Ethnea* is a collection of more than 9 million author name instances that are tagged one of 26 ENG classes based on the name's association with national-level geo-locations[10]. For example, "Wang, Wei" is classified as 'Chinese' as it is most frequently associated with organizations in China. However, *Ethnea* makes no distinctions based on any anthropological, cultural, or linguistic characteristics of authors. Instead it relies entirely on observations of names and geo-locations of their frequently associated institutions so an author named 'Wang, Wei' who was born in the U.S. and has never visited China would still be assigned a 'Chinese name' tag. We link all four labelled datasets to *Ethnea* and, if a matched name is found, assign that name's ENG tag to all its observed instances. If a queried name does not have a match in *Ethnea*, we search again using only the surname and assign the modal *Ethnea* ENG tag associated with it to all its instances. Table 1 summarizes the frequencies and ratios of ENG tags assigned by *Ethnea* to author name instances in each labeled dataset.

Table 1 shows the list of ENGs in each labeled data. Small-sized ENGs are excluded from analysis because most name instances in those ENGs tend to belong to a single author while a few instances referring to other author(s). When randomly split into training and test subsets for machine learning, these instances do not produce negative pairs at all.

Table 1: Summary of Ethnic Name Group (ENG) Frequencies in Labeled Data

Labeled Data (No. of Instances)							
KISTI (41,605)		**AMINER (7,528)**		**GESIS (29,965)**		**UM-IRIS (6,000)**	
ENG	**Ratio (%)**	**ENG**	**Ratio (%)**	**ENG**	**Ratio (%)**	**ENG**	**Ratio (%)**
CHINESE	50.1	CHINESE	64.2	CHINESE	56.2	CHINESE	16.67
ENGLISH	15.6	ENGLISH	17.0	GERMAN	14.8	ENGLISH	16.67
INDIAN	9.5	INDIAN	6.3	ENGLISH	7.0	GERMAN	16.67
KOREAN	8.0	GERMAN	5.6	INDIAN	4.3	HISPANIC	16.67
GERMAN	3.3	HISPANIC	3.1	HISPANIC	3.6	INDIAN	16.67
ISRAELI	2.2	**Sum**	**96.2**	KOREAN	3.5	KOREAN	16.67
ITALIAN	2.0	Excluded ENGs (3.8%): JAPANESE, NORDIC, KOREAN, ARAB		JAPANESE	2.8	**Sum**	**100.00**
HISPANIC	1.7			ITALIAN	1.7	Excluded ENGs (0%): None	
JAPANESE	1.1			ARAB	1.6		
DUTCH	1.0			FRENCH	1.3		
ARAB	0.9			**Sum**	**96.8**		
FRENCH	0.9			Excluded ENGs (3.2%): NORDIC, HUNGARIAN, DUTCH, VIETNAMESE, SLAV, ROMANIAN, GREEK, ISRAELI, TURKISH, INDONESIAN			
Sum	**96.3**						
Excluded ENGs (3.7%) : NORDIC, SLAV, GREEK, ROMANIAN, VIETNAMESE, NULL, AFRICAN, TURKISH, HUNGARIAN,							

[9] https://databank.illinois.edu/datasets/IDB-9087546

[10] 26 ethnicities include: African, Arab, Baltic, Caribbean, Chinese, Dutch, English, French, German, Greek, Hispanic, Hungarian, Indian, Indonesian, Israeli, Italian, Japanese, Korean, Mongolian, Nordic, Polynesian, Romanian, Slav, Thai, Turkish, and Vietnamese. In *Ethnea*, some name instances are assigned two ethnicities (e.g., "Jane Kim" → Korean-English) if the surname and forename of an author name are associated frequently with different ethnicities.

Chinese names represent the majority of ENGs in three labeled data. This is because these labeled data were created from computer science papers where Chinese researchers are particularly large contributors. In addition, as the three datasets were designed to collate challenging names to disambiguate, Chinese names that tend to be more ambiguous than other ENGs were over-sampled (Müller, Reitz, & Roy, 2017). In contrast, 6,000 instances in UM-IRIS are evenly distributed over six ENGs. For validation and reuse, these labeled data with ENG tags are publicly available[11]. Note that the original KISTI contains 41,673 name instances, whereas the ENG-tagged KISTI has 41,605 instances. Such discrepancy occurs because this paper uses the revised version of KISTI that corrects record errors and duplicates in the original data (J. Kim, 2018).

Machine Learning Process

Machine learning methods for AND can be divided into two groups: author assignment and author grouping (Ferreira et al., 2012). While the former aims to assign an author name instance to one of pre-disambiguated author name clusters, the latter aims to group all and only instances that belong to the same authors. This study takes the latter approach in evaluating the effect of ENG on AND. Specifically, author name instances in each labeled dataset are pairwise compared to assess whether a given instance pair of plausibly represents the same author (a match) or not (a nonmatch). Although some scholars take a further step to cluster pairwise comparisons (e.g., J. Kim & Kim, 2018; Levin, Krawczyk, Bethard, & Jurafsky, 2012; Louppe et al., 2016; Alan Filipe Santana et al., 2017), this study only evaluates disambiguation performance at a pair level (i.e., classification), following the practice of previous AND studies (e.g., Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004; Song et al., 2015; Treeratpituk & Giles, 2009; Vishnyakova, Rodriguez-Esteban, Ozol, & Rinaldi, 2016).

As the first machine learning step, author name instances in each labeled dataset are randomly divided into training (50%) and test (50%) subsets. Then, instances in each subset are put into blocks in which all member instances share the same full surname and first forename initial (e.g., 'Wang, W'). Only instances in the same block are compared for disambiguation. This blocking is typical in AND studies because it reduces computational complexity with only slight performance degradation (K. Kim, Sefid, & Giles, 2017; Torvik & Smalheiser, 2009). Next, instance pairs in the same block are compared to establish their similarity over four other data features: author name, coauthor name(s), paper title, and publication venue. To quantify how much a pair is similar over a feature, this study calculates the cosine similarity of Term (*n*-gram) Frequency for each feature (Han, Zha, & Giles, 2005; J. Kim & Kim, In print; Levin et al., 2012; Louppe et al., 2016; A. F. Santana, Gonçalves, Laender, & Ferreira, 2015; Treeratpituk & Giles, 2009). Specifically, the string of a feature is converted into an array of 2~4-grams (e.g., author name 'Wang, Wei' → 'wa|an|ng|gw|we|ei|wan|ang|ngw|gwe|wei|wang|angw|ngwe|gwei'). After the conversion, two *n*-gram arrays of an instance pair are compared to produce a cosine similarity score for the feature.

Besides the four basic features, ENGs are used as a feature set for ENG-aware disambiguation. For this, especially, an instance pair's ENG is encoded into a binary value (i.e., one-hot encoding) for a pre-defined set of ENGs[12]. For example, in AMINER, a pair of name instances ('Wang, Wei' and 'Wang, W.') is assigned either 'Yes' or 'No' for each of five ethnicities – Chinese ('Yes'), English ('No'), Indian

[11] Download link TBA

[12] As only instances in the same block in which they share at least the same full name and first forename initial are compared, all the pairs in the block have the same ethnicity tag.

(‘No’), German (‘No’), and Hispanic (‘No’) – as shown in Table 1. Table 2 shows examples of the cosine similarity scores calculated over four features and ENG encoding results for instance pairs.

Table 2: A Mock-Up Example of Cosine Similarity Scores for Instance Pairs over Four Features and ENG Encoding

Pairs	Feature								Label
	Author Name	Coauthor Name	Paper Title	Pub. Venue	ENG 1	ENG 2	ENG 3	…	
Pair 1	0.97	0.89	0.67	0.12	Yes	No	No	**…**	Match
Pair 2	1.00	0.24	0.46	0.00	No	Yes	No	**…**	Nonmatch
Pair 3	0.65	0.07	0.00	0.80	No	No	Yes	**…**	Nonmatch
Pair 4	0.58	0.08	0.00	0.00	Yes	No	No	**…**	Nonmatch

We focus on four algorithms – Gradient Boosting, Logistic Regression, Naïve Bayes, and Random Forest – for supervised machine learning that have been widely used as baselines or best performing methods in AND studies (e.g., Han et al., 2004; J. Kim & Kim, In print; K. Kim, Sefid, Weinberg, & Giles, 2018; Louppe et al., 2016; Song et al., 2015; Torvik & Smalheiser, 2009; Treeratpituk & Giles, 2009; Vishnyakova et al., 2016; J. Wang et al., 2012). In the first scenario, they are trained on the list of similarity scores and labels, as shown in Table 2 to learn relative weights for features and an absolute weight or threshold for instance pairs to be disambiguated without considering ENGs (→ ENG-ignorant learning). In the second scenario, the same algorithms are trained on the list of similarity scores, ENGs, and labels (→ ENG-aware disambiguation). Here, the ethnic name partition adds more features (dimensions) to each instance pair’s feature set, allowing algorithms to combine the similarities of the expanded features. The machine learning procedure is implemented using the python Scikit-learn package. For Gradient Boosting, 500 estimators are used with max depth=9 and learning rate = 0.125. For Logistic Regression, L2 Regularization with class weight = 1 is used. Gaussian Naïve Bayes with maximum likelihood estimator is used for Naïve Bayes. For Random Forest, 500 trees are used after a grid search.

Trained algorithmic models are applied to the instance pairs in test subsets in which the cosine similarity is calculated for the four basic features and, in the second scenario, ethnicities are encoded in the same fashion but explicitly include ENG information. As in Table 2, an algorithmic model receives a set of feature similarity scores and, if ENG-aware disambiguation is conducted, a list of encoded ENGs for an instance pair to output a binary classification decision (match or nonmatch). Once trained, each algorithm produces a single score that predicts the probability of an instance pair being negative (nonmatch). If the predicted probability is above a certain threshold (> 0.5), the pair is decided to be a nonmatch, whereas if below the threshold, a match.

Performance Evaluation

We evaluate each algorithm’s classification results on reserved test subsets of each labeled dataset by calculating precision and recall for positive (P; match) and negative (N; nonmatch) pairs respectively. In addition, we calculate the F1 score as a harmonic mean of precision and recall.

Specifically, precision for positive pairs (Prec-Pos) measures how many predicted match pairs are correct ones (true positives; TP) over the total number of predicted match pairs that may contain correct match pairs (true positives; TP) and incorrect match pairs (false positives; FP). In contrast, recall for positive pairs (Rec-Pos) measures the ratio of correct match pairs (true positives; TP) over the total number of true match pairs that may be predicted correctly as match pairs (true positives; TP) or incorrectly as nonmatch pairs (false negatives; FN).

$$Precision\ Positive\ (PrecPos) = \frac{Number\ of\ Correctly\ Predicted\ Match}{Number\ of\ Predicted\ Match} = \frac{TP}{(TP + FP)} \quad (1)$$

$$Recall\ Positive\ (RecPos) = \frac{Number\ of\ Correctly\ Predicted\ Match}{Number\ of\ True\ Match} = \frac{TP}{(TP + FN)} \quad (2)$$

$$F1\ Positive = \frac{2 \times PrecPos \times RecPos}{PrecPos\ +\ RecPos} \quad (3)$$

Likewise, precision for negative pairs (Prec-Neg) measures how many predicted nonmatch pairs are correct ones (true negatives; TN) over the total number of predicted nonmatch pairs that may contain correct nonmatch pairs (true negatives; TN) and incorrect nonmatch pairs (false negatives; FN). In contrast, recall for negative pairs (Rec-Neg) measures the ratio of correct nonmatch pairs (true negatives; TN) over the total number of true nonmatch pairs that may be predicted correctly as nonmatch pairs (true negatives; TN) or incorrectly as match pairs (false positives; FP).

$$Precision\ Negative\ (PrecNeg) = \frac{Number\ of\ Correctly\ Predicted\ Nonmatch}{Number\ of\ Predicted\ Nonmatch} = \frac{TN}{(TN + FN)} \quad (4)$$

$$Recall\ Negative\ (RecNeg) = \frac{Number\ of\ Correctly\ Predicted\ Nonmatch}{Number\ of\ True\ Nonmatch} = \frac{TN}{(TN + FP)} \quad (5)$$

$$F1\ Negative = \frac{2 \times PrecNeg \times RecNeg}{PrecNeg\ +\ RecNeg} \quad (6)$$

The metrics are calculated on the entire set of test results for each labeled dataset. We separately calculate performance measures for difference ENGs rather than averaging them across multiple ethnicity groups.

Results

Cross-Data Performance Evaluation

Figure 1 shows disambiguation results on KISTI, reporting precision and recall *before* and *after* ENG-aware disambiguation by four algorithms – Gradient Boosting (GB), Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF). Figure 1a shows that when ENGs are included as features, the algorithms tend to produce better precision in the prediction of positive (match) pairs than when they are not considered. This is shown by black bars ('After') being higher than stripped bars ('Before') in Figure 1a. This observation indicates that ethnic name partitioning helps algorithms increase the ratio of TP among predicted positive pairs (= TP + FP). This can be confirmed by checking the numbers of true and false positive pairs in Table 3. For example, when trained only on the four basic (non-ENG) features, LR predicts that 76,201 (= TP + FP = 55,998 + 20,203) pairs refer to the same authors (match) and 73.49% of the predictions are right (= TP/(TP + FP)). After trained on the same but ENG-tagged data, however, it predicts 170,432 pairs to be match sets, increasing its prediction accuracy this time to 77.08%.

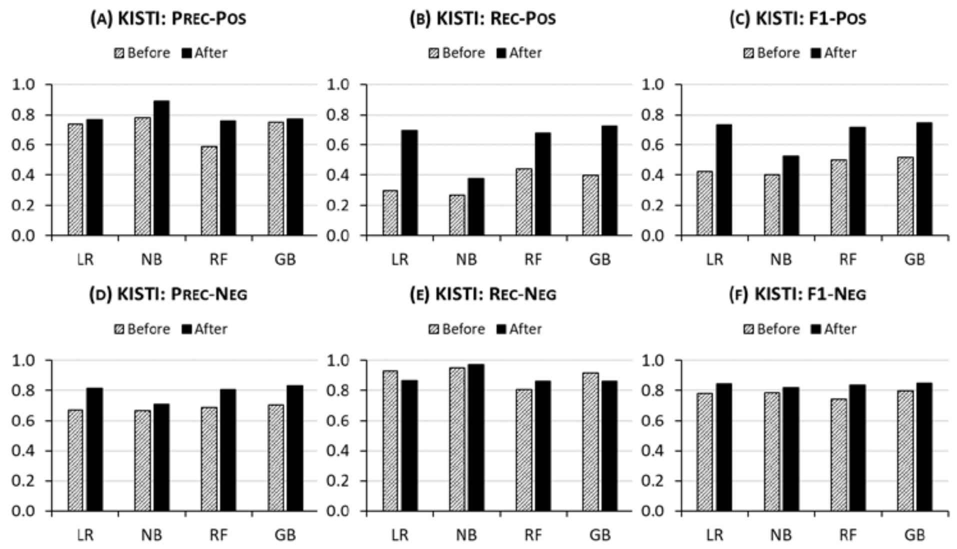


Figure 1: Disambiguation Performances 'Before' Versus 'After' ENG-Aware Disambiguation on KISTI

Table 3: Numbers of Correctly or Incorrectly Predicted Pairs for Positive and Negative Pairs by Four Algorithms on KISTI Test Data

Algorithm	ENGs Considered	No. of Pairs (P: Positive) (N: Negative)	TP	FN	FP	TN
LR	Before	483,029 (P: 189,375) (N: 293,654)	55,998	133,377	20,203	273,451
	After		131,364	58,011	39,068	254,586
NB	Before		50,782	138,593	14,267	279,387
	After		71,020	118,355	8,775	284,879
RF	Before		82,727	106,648	57,532	236,122
	After		128,105	61,270	40,479	253,175
GB	Before		75,195	114,180	24,725	268,929
	After		136,859	52,516	39,888	253,766

Ethnic name partitioning also reduces the number of falsely predicted nonmatch cases (FN), increasing recall in Figure 1b. Performance gains by ENG-aware disambiguation are more pronounced for recall than for precision, as evidenced by larger differences between 'Before' and 'After' bars for recall (Figure 1b) than those for precision (Figure 1a). In other words, ENG aware disambiguation across four common algorithms appears to reduce false negative predictions more than true positive predictions, potentially providing better performance for applications (such as network analysis) that are particularly sensitive to biases due to erroneous "lumping" of name instances that actually refer to different individuals. The improvements in precision and recall together increase the F1 scores by ENG-aware disambiguation (Figure 1c).

ENG-aware disambiguation also does a better job of accurately predicting non-match (negative pair) cases. The 'After' bars are taller than those of 'Before' in Figure 1d. Unlike the positive pair prediction in which ENG-aware disambiguation works in favor of both precision and recall by all algorithms, however, the performance gains in precision for negative pairs come with slightly decreased recall by GB and LR in Figure 1e. This means that while disambiguation models by GB and LR trained on ENG-added features are good at increasing the numbers of true nonmatch pairs among predicted nonmatch pairs (= TN +FN), they incorrectly predict that true nonmatch pairs match (FP predictions) more frequently than when they are trained on the four basic features alone. Reduced recall for negative pair prediction is, however, offset by increased precision, leading to the F1 scores by ENG-aware disambiguation being better than those by ENG-blind one in Figure 1f. Meanwhile, NB and RF still obtain improvements in both precision and recall as well as F1.

Algorithmic performances are also enhanced by ENG-aware disambiguation on AMINER, GESIS, and UM-IRIS. Figure 2 ~ 4 report that the algorithms trained on ENG-tagged data perform better than those trained only on the basic features across almost all metrics for both positive and negative pairs. NB models prove the exception, producing worse results in recall for positive pairs and in precision for negative pairs after ethnic name partition. However, this degraded performance is offset by increased precision for positive pairs and increased recall for negative pairs, respectively, so the overall performance metric (F1), which equally weights precision and recall, indicates an overall improvement due to the inclusion of ENG features.

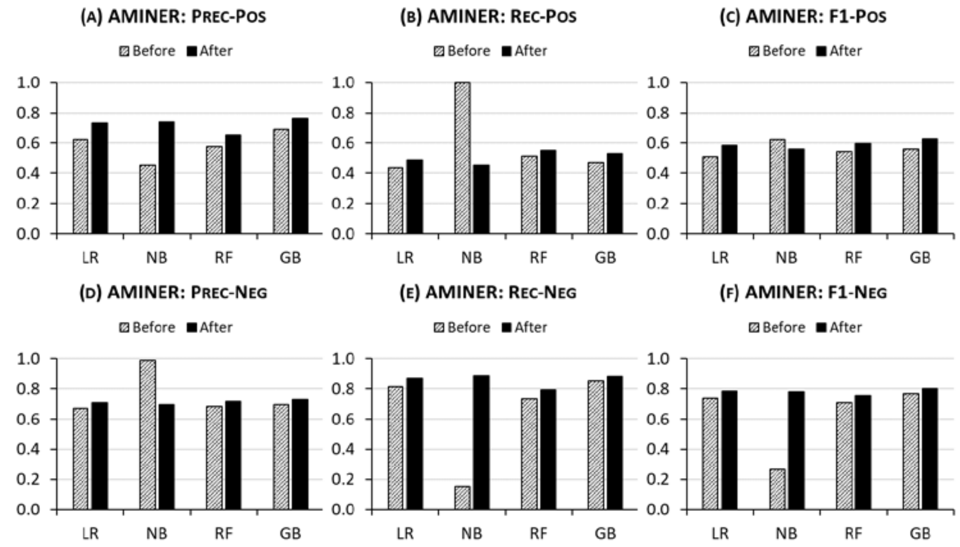


Figure 2: Disambiguation Performances 'Before' Versus 'After' ENG-Aware Disambiguation on AMINER

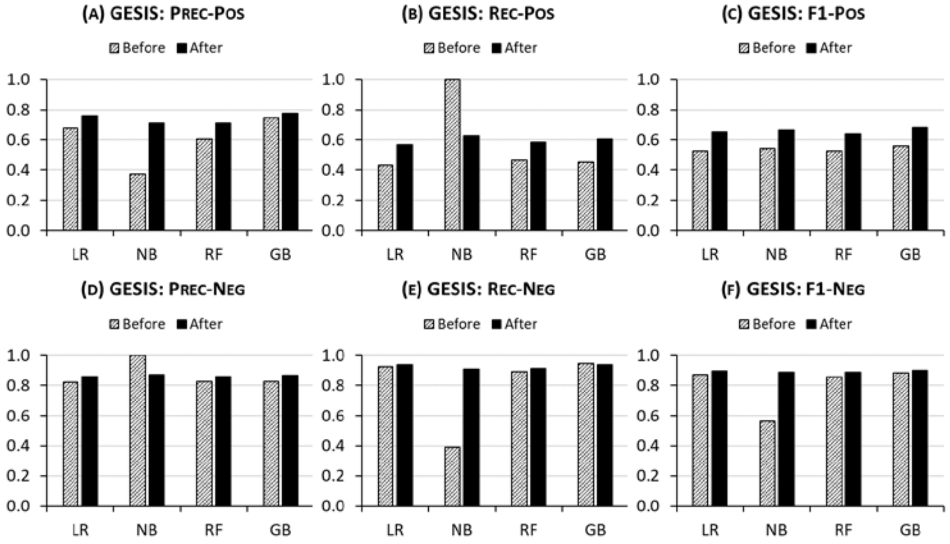


Figure 3: Disambiguation Performances 'Before' Versus 'After' ENG-Aware Disambiguation on GESIS

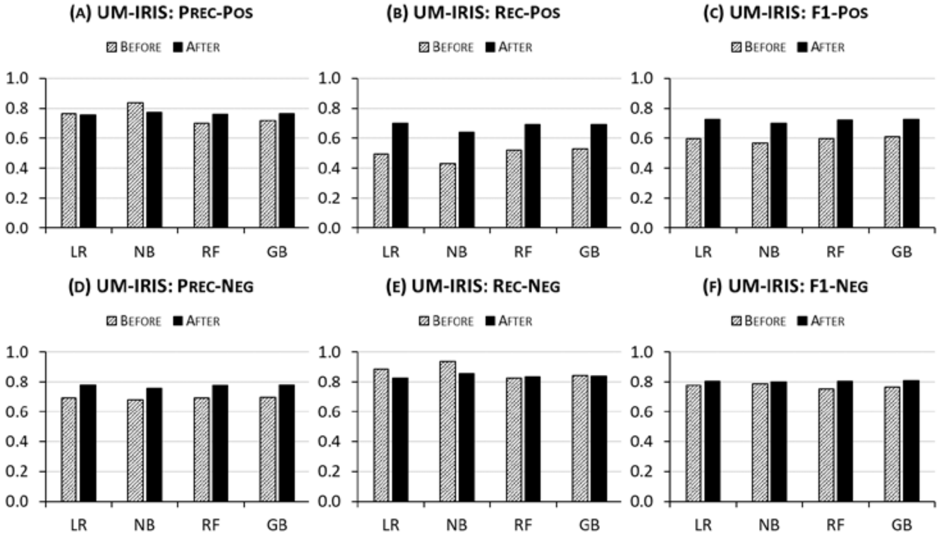


Figure 4: Disambiguation Performances 'Before' Versus 'After' ENG-Aware Disambiguation on UM-IRIS

Performance Evaluation per ENG

ENG-aware disambiguation produces substantial improvements in both precision and recall for predicting match and nonmatch instance pairs in different labeled datasets. But are those improvements uniform across different ENGs? If not, a more nuanced approach to model evaluation may be necessary. To answer this question, we compare performance changes due to ENG-aware disambiguation within ENG groups. For this, precision, recall, and F1 scores for positive and negative pairs predicted by four algorithms are calculated separately for instance pairs that belong to the same ENG in each of four labeled data: 4 algorithms × 4 data = 16 evaluations. Presenting all the results at the same time would consume too much space in this paper. So, we present random forest (RF) predictions on the GESIS dataset as an illustration for the purposes of this discussion. Reports of other algorithms and data are presented in a supplementary document attached to this paper.

Figure 5 shows the by ENG performance metrics for the RF algorithm trained on GESIS with and without ENG-aware disambiguation. The ENG-aware disambiguation leads to better precision (positive pairs; Figure 5A) and recall (negative pairs; Figure 5E) for Chinese names but worse precision (positive pairs) and recall (negative pairs) for other ethnicities. In contrast, name disambiguation for Chinese names results in lower recall (positive pairs; Figure 4B) and precision (negative pairs; Figure 4D) than those for other ENGs. Similar patterns are observed for other algorithms tested on GESIS (see Figure S5~S8 in Supplementary Material). This suggests that the effect of ENG-aware disambiguation occurs in different ways for different ENGs. Thus, its application can be beneficial in some instances but detrimental in others. Variations in the effects of ENG-aware disambiguation on precision and recall for positive and negative pair prediction across ethnicity groups suggest that care must be taken to design disambiguation strategies that fit particular analytic or empirical needs.

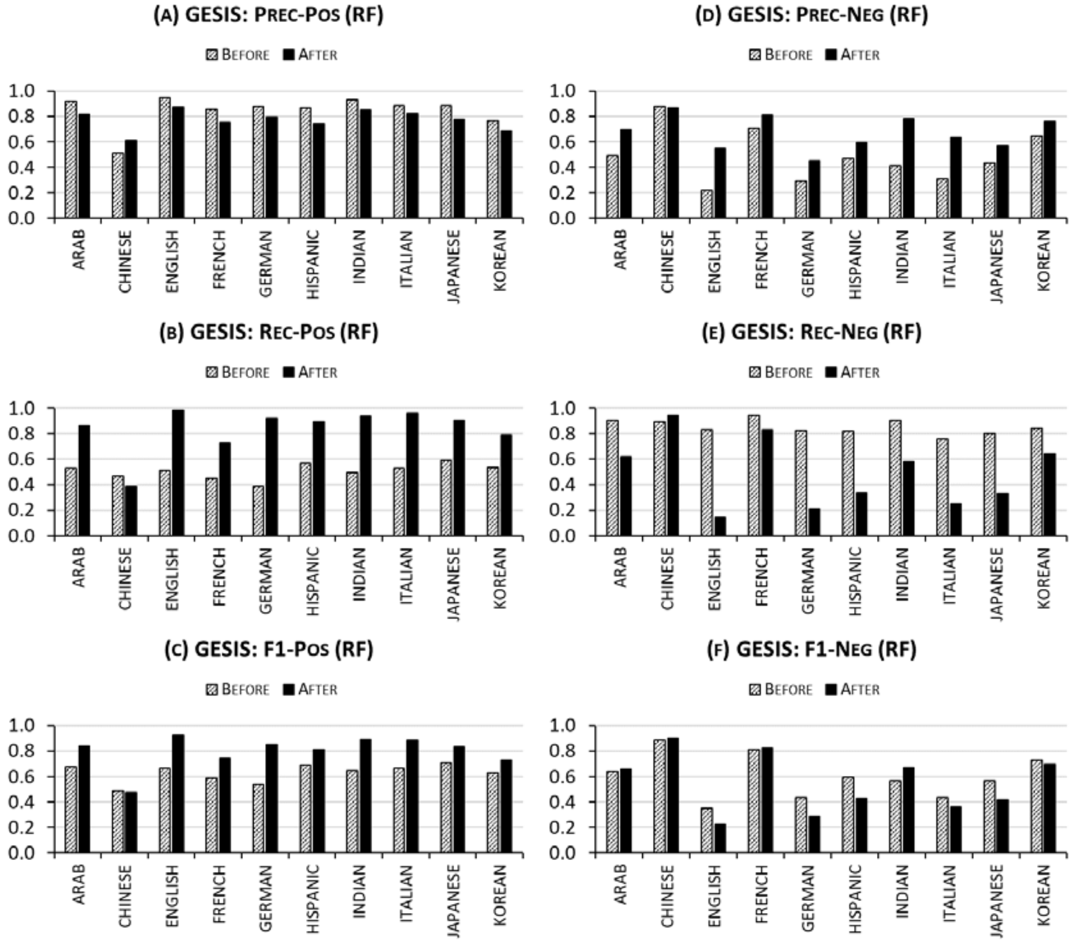


Figure 5: Disambiguation Performances per ENG 'Before' Versus 'After' ENG-Aware Disambiguation by Random Forest on GESIS

These observations can be explained as follows. ENGs have different distributions of similarity scores over the four basic (non-ethnicity) features we use. Figure 6 presents the feature similarity score distributions per ENG for positive (left) and negative (right) pairs in the GESIS test data. Training and test subsets show similar distributions in each labeled dataset. For visual simplicity, a score is rounded up into nearest bins with intervals of 0.1 on *x*-axis and the ratios of the numbers of scores in the same bin over all scores are plotted on *y*-axis. A solid red line represents the distribution of all instance pairs regardless of ENG.

In Figure 6, each ENG has different distributions of, for example, 'COAUTHOR' similarity scores for both positive and negative pairs (Figure 6C and 6D). So, the four algorithms come to use different 'coauthor' similarity score distributions in ENG-aware disambiguation. Such heterogeneous distributions also occur for other features but with different variations of differences. For example, 'VENUE' distributions in Figure 6G and 6H differ less across ENGs than do 'COAUTHOR' distributions. Because ENG-aware disambiguation allows training and testing on different feature similarity score distributions for each ENG, the algorithms combine features using different weightings for each ethnicity, producing different predictions for name pairs with the same feature similarity scores but different ENG tags. In other words, this method takes into account the likelihood that researchers in different ENGs organize their scientific work differently, favoring distinct co-authorship and publication venue patterns. This also occurs in disambiguation of other labeled data, whose feature similarity score distributions are reported in Figure S17 ~ S20 in Supplementary Material.

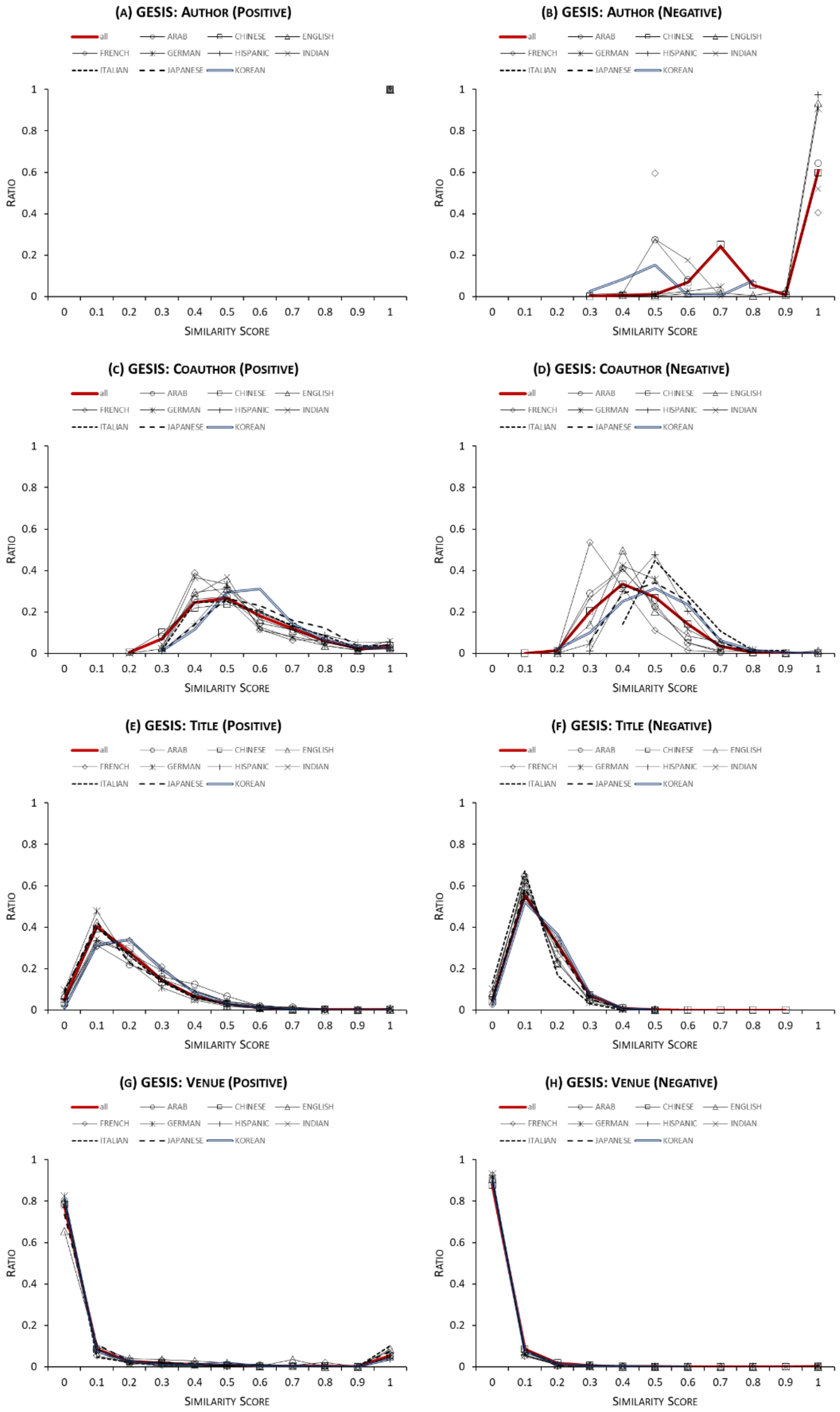


Figure 6: Feature Similarity Score Distributions per ENG for Positive and Negative Pairs in GESIS Test Data

Figure 5 also shows that some ENGs manifest substantial improvements in recall for positive pairs (Figure 5B) but degraded recall for negative pairs (Figure 5E). This might be explained in two ways. In our 'before' (ENG-unaware) case, algorithms combine features to produce per-feature weights for positive pairs based on feature similarity scores aggregated across multiple ENGs that can have very different feature distributions. Such aggregated distributions cannot effectively capture the single match patterns specific to each ENG, which seem to lead models to falsely predict positive pairs as negative ones (FN), reducing the recall for positive pairs. Conversely, increased recall for positive pairs after ENG-aware disambiguation means that the algorithms trained and tested on ENG-tagged data successfully produce per-feature weights optimized to each ENG, thus making better predictions that push up the recall scores for many ENGs.

Second, decreased negative pair recall after ENG-aware disambiguation means that the algorithms trained and tested on ENG-tagged data fail to produce proper per-feature weights for accurately predicting nonmatch for known negative pairs. When the algorithms are trained only on the four basic (non-ethnicity) features, they do a better job of predicting nonmatch pairs based on aggregated feature similarity distributions that are invariant across particular ENGs. In other words, feature distributions aggregated across ENGs appear to be more effective for predicting negative case pairs while ENG-aware disambiguation techniques more accurately capture positive pairs.

These observations imply that disambiguation models for positive pair prediction would be improved by ENG-aware procedures, while nonmatch patterns for negative pair prediction can aggregate across ENGs (J. Kim & Kim, 2018). Table 4 shows that in the GESIS training data, each ENG has different sizes of positive and negative (pairwise) pairs. CHINESE name instances produce the largest numbers of positive (≈ 146K) and negative pairs (≈ 551K), while ITALIAN name instances generate around a few thousand positive and a few hundred negative pairs. In other training data, CHINESE pairs constitute substantially large proportions (KISTI: 71.28 % and AMINER: 91.26 %) or over one-third (UM-IRIS: 37.54 %) of all negative pairs, while other ENG pairs make up small or less-than-expected (approximately 17% per ENG in UM-IRIS) proportions. In contrast, the numbers of positive pairs are less concentrated (GESIS, KISTI, and AMINER) or more evenly distributed (UM-IRIS) for positive pairs than those for negative pairs.

Table 4: Distributions of Positive and Negative Name Instance Pairs per ENG in GESIS Training Data

ENG	**Positive Pairs**	**Ratios**	**Negative Pairs**	**Ratios**
ARAB	1,559	0.66	1,203	0.21
CHINESE	145,969	62.22	551,150	93.96
ENGLISH	16,020	6.83	2,147	0.37
FRENCH	1,948	0.83	3,118	0.53
GERMAN	37,373	15.93	13,092	2.23
HISPANIC	5,887	2.51	1,907	0.33
INDIAN	6,708	2.86	2,654	0.45
ITALIAN	3,489	1.49	686	0.12
JAPANESE	8,853	3.77	2,608	0.44
KOREAN	6,792	2.90	7,994	1.36
Total	234,598	100	586,559	100

Table 5: Distributions of Positive and Negative Name Instance Pairs per ENG in KISTI Training Data

ENG	Positive Pairs	Ratios	Negative Pairs	Ratios
ARAB	6,015	3.21	37	0.01
CHINESE	53,519	28.55	209,807	71.28
DUTCH	5,448	2.91	84	0.03
ENGLISH	45,054	24.04	25,137	8.54
FRENCH	2,686	1.43	625	0.21
GERMAN	11,065	5.90	4,114	1.40
HISPANIC	3,729	1.99	1,680	0.57
INDIAN	33,742	18.00	19,043	6.47
ISRAELI	9,699	5.17	460	0.16
ITALIAN	9,598	5.12	283	0.10
JAPANESE	1,588	0.85	580	0.20
KOREAN	5,284	2.82	32,474	11.03
Total	187,427	100	294,324	100

Table 6: Distributions of Positive and Negative Name Instance Pairs per ENG in AMINER Training Data

ENG	Positive Pairs	Ratio (%)	Negative Pairs	Ratio (%)
CHINESE	38,958	61.88	76,781	91.26
ENGLISH	13,758	21.85	2,414	2.87
GERMAN	2,603	4.13	748	0.89
HISPANIC	1,933	3.07	380	0.45
INDIAN	5,701	9.06	3,815	4.53
Total	62,953	100	84,138	100

Table 7: Distributions of Positive and Negative Name Instance Pairs per ENG in UM-IRIS Training Data

ENG	Positive Pairs	Ratio (%)	Negative Pairs	Ratio (%)
CHINESE	1,517	10.60	7,261	37.54
ENGLISH	2,185	15.26	459	2.37
GERMAN	3,673	25.66	1,917	9.91
HISPANIC	2,416	16.88	970	5.01
INDIAN	2,080	14.53	3,803	19.66
KOREAN	2,445	17.08	4,933	25.50
Total	14,316	100	19,343	100

As noted above for Figure 5 and observed in other labeled data (see Figure S1 ~ S16 in Supplementary Material), the algorithms work better in finding more true negative pairs even for non-CHINESE name pairs when they are trained on data in which ethnic name partitioning is not performed (‘BEFORE’) and, thus, negative pairs are dominated by CHINESE ones as shown in Table 4~6. This implies that the nonmatch patterns in CHINESE name pairs are applicable to predicting nonmatch pairs for other ENGs. In contrast, during ENG-aware disambiguation, the algorithms come to rely on the small-size negative pairs that may skew or distort true nonmatch patterns for some ENGs. This seems to result in the decreased recall in predicting negative pairs (i.e., many true negatives classified as false positives, which reduces precision for positive pair prediction), while increasing slightly precision in predicting negative pairs.

Despite the aforementioned conflicting changes in precision and recall per ENG, the overall performance by the four algorithms on the whole test set are shown in Figure 1~4 to substantially increase across the four labeled data after ethnic name partitioning is included in machine learning. One reason would be that performance gains outweigh losses at each ENG level overall. Another reason would be that especially for KISTI, AMINER, and GESIS, the improved performances in disambiguating CHINESE that constitute the majority of name instances may affect the overall evaluation results. As shown by the case of UM-IRIS in which ENG sizes are controlled to be equal, however, the overall performance improvements can be observed for all the ENGs by ENG-aware disambiguation. As such, this study illustrates that the ethnic name partition can be truly effective in improving disambiguation performances.

Discussion

These results suggest that AND tasks may produce better results by using ethnic name partition in machine learning. Considering that adding more features can improve generally machine learning performances, the enhanced disambiguation performances by ENG partitioning might not be a surprise. With that said, the real contribution of this study would be that it demonstrates many machine learning based disambiguation models have a potential to be improved by introducing ethnic name grouping into ambiguous data without additional collection of feature information.

To fully realize this potential, however, a few issues need to be addressed. First, ENG tagging can be a non-trivial task that requires an intricate algorithmic technique itself. Thanks to the ENG classification system developed and publicly shared by Torvik and Agarwal (2016), this study could assign ENGs to the names in four labeled data. Although *Ethnea* was constructed based on more than 9 million author name instances in PubMed, the world largest biomedicine library, it is unknown how well it can help us tag ENGs to names in other fields. Ideally, *Ethnea* may be updated regularly to reflect new author names entering bibliographic data in various fields. Practically, further research may be focused on finding out a set of ENGs that are most influential in improving disambiguation results and thus simplifying ENG tagging for author name disambiguation (e.g., CHINESE vs Non-CHINESE).

Second, the findings of this study were based on three labeled data (KISTI, AMINER, and GESIS) in which CHINESE names are dominant and the overall performance improvements were heavily affected by those for CHINESE name instances. To overcome such an imbalance of instance distribution in labeled data, a new labeled data (UM-IRIS) were created in a way that six ENGs have the same amount of ambiguous name instances. Disambiguation results from the new labeled data were in line with those from other three labeled data. In addition, all ENGs including CHINESE were able to obtain gains in disambiguation performances. But all these findings were obtained from small-sized labeled data, whether they are biased or controlled for ENG sizes. So, it is still unknown whether such improvements are achievable in author name disambiguation for large-scale bibliographic data in which ENG composition may be quite different from those in the labeled data used in this paper.

Another issue would be that there can be other features than the four used in this study that can lead ethnic name partition to different AND performances. For example, English authors may appear in publication records that are more complete in affiliation information and use more diverse title terms. Meanwhile, Chinese authors may tend to work with coauthors who have similar names in same institutions. Various features need to be explored to study further the impact of ethnic name partition on AND.

Fourth, ENG-aware disambiguation may be beneficial for positive pair prediction but not so much for negative pair prediction. This was illustrated in Figure 5 above and Figure S1 ~ S16 in Supplementary Material by the dramatically decreased recall in negative pair predictions for many ENGs. It was contrasted with the substantial increase of precision in positive pair prediction for those ENGs. This study speculates that by ethnic name partitioning, classifiers become stricter for CHINESE pairs while relaxed for other ENG pairs. In other words, a pairs of CHINESE instances that would be classified as 'match' before partitioning are classified as 'nonmatch' after partitioning (PREC-POS↑; REC-POS↓), while 'nonmatch' pairs of other ENG instances as matched ones (PREC-POS↓; REC-POS↑). This might be because while some CHINESE pairs sharing coauthor names, venue names, or title words refer to different authors, other ethnic names sharing the features are more likely to represent the same authors (see Figure 6 B, D, F, and H in which Chinese name pair share is denoted in square). During training, such different similarity patterns are mixed up before partitioning but distinguished after it. Another conjecture is that due to the relaxed classification after partitioning, true negative pairs of other-than-CHINESE ENGs are falsely classified as false positive pairs (PREC-POS↓; REC-NEG↓). As the sizes of negative pairs in most non-CHINESE ENGs are smaller than positive pairs (see Table 4~7), misclassified negative pairs have larger impacts on REC-NEG than on PREC-POS across the ENGs. But these conjectures are based on the observations on labeled data in which Chinese name instances are prevalent. Using an additional labeled data with controlled ENG sizes, however, the conjecture has been confirmed. But only six ENGs in a small dataset were considered for analysis. More ENGs need to be investigated to check if this conjecture holds good under the different combinations of ENGs.

Conclusion and Discussion

This study evaluated the effects ethnic name partitioning has on author name disambiguation (AND) using machine learning methods. For this, author name instances in four labeled datasets were disambiguated under two scenarios. First, similarity scores of instance pairs over four basic features – author name, coauthor names, paper title, and publication venue – were used to train and test disambiguation algorithms. Second, in addition to the basic features, ethnic name group (ENGs) were tagged to name instances to allow algorithms to build models that are optimized to each ENG. Comparisons of disambiguation performances before and after ENG-aware disambiguation showed that using ethnic name partition can substantially improve algorithmic performances. Such performance improvements occurred across all ENGs, although performance gains and losses at each ENG level were observed in different ways depending on the types of measures – precision or recall – and target classifications – positive (match) or negative (nonmatch) pairs.

As detailed in the discussion above, ethnic name partition requires further research for us to understand better its impact on author name disambiguation and apply it to disambiguation tasks for digital libraries that are struggling with authority control over fast-growing ambiguous author names. This study is expected to motivate scholars and practitioners to study toward that direction by demonstrating the potential of ENG-aware disambiguation in improving disambiguation performances.

Acknowledgements

This paper was supported by grants from the National Science Foundation (#1917663, #1760609, and #1561687). The authors would like to thank anonymous reviewers for their comments to improve the paper.

References

Chin, A. W.-S., Juan, Y.-C., Zhuang, Y., Wu, F., Tung, H.-Y., Yu, T., . . . et al. (2013). *Effective string processing and matching for author disambiguation*. Paper presented at the Proceedings of the 2013 KDD Cup 2013 Workshop, Chicago, Illinois. https://doi.org/10.1145/2517288.2517295

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. F. (2010). An Unsupervised Heuristic-Based Hierarchical Method for Name Disambiguation in Bibliographic Citations. *Journal of the American Society for Information Science and Technology, 61*(9), 1853-1870. doi:10.1002/asi.21363

Deville, P., Wang, D. S., Sinatra, R., Song, C. M., Blondel, V. D., & Barabási, A.-L. (2014). Career on the Move: Geography, Stratification, and Scientific Impact. *Scientific Reports, 4*. doi:10.1038/srep04770

Fegley, B. D., & Torvik, V. I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLOS ONE, 8*(7). doi:10.1371/journal.pone.0070299

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. *Sigmod Record, 41*(2), 15-26.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. F. (2014). Self-Training Author Name Disambiguation for Information Scarce Scenarios. *Journal of the Association for Information Science and Technology, 65*(6), 1257-1278. doi:10.1002/asi.22992

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). *Two Supervised Learning Approaches for Name Disambiguation in Author Citations.* Paper presented at the Jcdl 2004: Proceedings of the Fourth Acm/Ieee Joint Conference on Digital Libraries, Tucson, Arizona.

Han, H., Zha, H. Y., & Giles, C. L. (2005). Name disambiguation in author citations using a K-way spectral clustering method. *Proceedings of the 5th Acm/Ieee Joint Conference on Digital Libraries, Proceedings*, 334-343. doi:Doi 10.1145/1065385.1065462

Harzing, A. W. (2015). Health warning: might contain multiple personalities-the problem of homonyms in Thomson Reuters Essential Science Indicators. *Scientometrics, 105*(3), 2259-2270. doi:10.1007/s11192-015-1699-y

Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review, 32*, e22. doi:10.1017/S0269888917000182

Hussain, I., & Asghar, S. (2018). DISC: Disambiguating homonyms using graph structural clustering. *Journal of Information Science, 44*(6), 830-847. doi:10.1177/0165551518761011

Islamaj Dogan, R., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database : the journal of biological databases and curation, 2009*, bap018-bap018. doi:10.1093/database/bap018

Kang, I. S., Kim, P., Lee, S., Jung, H., & You, B. J. (2011). Construction of a Large-Scale Test Set for Author Disambiguation. *Information Processing & Management, 47*(3), 452-465. doi:10.1016/j.ipm.2010.10.001

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics, 103*(3), 1061-1071. doi:10.1007/s11192-015-1580-z

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: A case of DBLP. *Scientometrics, 116*(3), 1867-1886. doi:10.1007/s11192-018-2824-5

Kim, J., & Diesner, J. (2015). The effect of data pre-processing on understanding the evolution of collaboration networks. *Journal of Informetrics, 9*(1), 226-236. doi:10.1016/j.joi.2015.01.002

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology, 67*(6), 1446-1461. doi:10.1002/asi.23489

Kim, J., & Owen-Smith, J. (In print). ORCID-linked labeled data for evaluating author name disambiguation at scale. Scientometrics. doi:10.1007/s11192-020-03826-6

Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics, 117*(1), 511-526. doi:10.1007/s11192-018-2865-9

Kim, J., & Kim, J. (2020). Effect of Forename String on Author Name Disambiguation. *Journal of the Association for Information Science and Technology, 71*(7), 839-855. doi:10.1002/asi.24298

Kim, K., Sefid, A., & Giles, C. L. (2017). Scaling Author Name Disambiguation with CNF Blocking. *arXiv preprint arXiv:1709.09657*.

Kim, K., Sefid, A., Weinberg, B. A., & Giles, C. L. (2018). *A Web Service for Author Name Disambiguation in Scholarly Databases.* Paper presented at the 2018 IEEE International Conference on Web Services (ICWS).

Lerchenmueller, M. J., & Sorenson, O. (2016). Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists. *PLOS ONE, 11*(7), e0158731. doi:10.1371/journal.pone.0158731

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-Based Bootstrapping for Large-Scale Author Disambiguation. *Journal of the American Society for Information Science and Technology, 63*(5), 1030-1047. doi:10.1002/asi.22621

Ley, M. (2009). DBLP: some lessons learned. *Proceedings of the VLDB Endowment, 2*(2), 1493-1500.

Louppe, G., Al-Natsheh, H. T., Susik, M., & Maguire, E. J. (2016). ENG Sensitive Author Disambiguation Using Semi-supervised Learning. *Knowledge Engineering and Semantic Web, Kesw 2016, 649*, 272-287. doi:10.1007/978-3-319-45880-9_21

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics, 7*(4), 767-773. doi:10.1016/j.joi.2013.06.006

Momeni, F., & Mayr, P. (2016). *Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names.* Paper presented at the 20th international Conference on Theory and Practice of Digital Libraries (TPDL 2016), Hannover, Germany.

Müller, M. C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics, 111*(3), 1467-1500. doi:10.1007/s11192-017-2363-5

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2015). On the Combination of Domain-Specific Heuristics for Author Name Disambiguation: The Nearest Cluster Method. *International Journal on Digital Libraries, 16*(3-4), 229-246. doi:10.1007/s00799-015-0158-y

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2017). Incremental author name disambiguation by exploiting domain-specific heuristics. *Journal of the Association for Information Science and Technology, 68*(4), 931-945. doi:10.1002/asi.23726

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2019). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*. doi:10.1177/0165551519888605

Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics, 107*(3), 1283-1298. doi:10.1007/s11192-016-1892-7

Shin, D., Kim, T., Choi, J., & Kim, J. (2014). Author Name Disambiguation Using a Graph Model with Node Splitting and Merging Based on Bibliographic Information. *Scientometrics, 100*(1), 15-50. doi:10.1007/s11192-014-1289-4

Smalheiser, N. R., & Torvik, V. I. (2009). Author Name Disambiguation. *Annual Review of Information Science and Technology, 43*, 287-313.

Song, M., Kim, E. H. J., & Kim, H. J. (2015). Exploring Author Name Disambiguation on PubMed-Scale. *Journal of Informetrics, 9*(4), 924-941. doi:10.1016/j.joi.2015.08.004

Strotmann, A., & Zhao, D. Z. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology, 63*(9), 1820-1833. doi:10.1002/asi.22695

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). *ArnetMiner: extraction and mining of academic social networks*. Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA.

Torvik, V. I., & Agarwal, S. (2016). *Ethnea: An instance-based ENG classifier based on geo-coded author names in a large-scale bibliographic database*. Paper presented at the Library of Congress International Symposium on Science of Science, Washington D.C. http://hdl.handle.net/2142/88927

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *Acm Transactions on Knowledge Discovery from Data, 3*(3). doi:10.1145/1552303.1552304

Treeratpituk, P., & Giles, C. L. (2009). Disambiguating Authors in Academic Publications using Random Forests. *JCDL 2009: Proceedings of the 2009 Acm/Ieee Joint Conference on Digital Libraries*, 39-48.

Vishnyakova, D., Rodriguez-Esteban, R., Ozol, K., & Rinaldi, F. (2016, dec). *Author Name Disambiguation in MEDLINE Based on Journal Descriptors and Semantic Types.* Paper presented at the Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), Osaka, Japan.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics, 93*(2), 391-411. doi:10.1007/s11192-012-0681-1

Wang, X., Tang, J., Cheng, H., & Yu, P. S. (2011). *ADANA: Active Name Disambiguation*. Paper presented at the 2011 IEEE 11th International Conference on Data Mining.

Wu, H., Li, B., Pei, Y. J., & He, J. (2014). Unsupervised author disambiguation using Dempster-Shafer theory. *Scientometrics, 101*(3), 1955-1972. doi:10.1007/s11192-014-1283-x

Wu, J., & Ding, X. H. (2013). Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics, 96*(3), 683-697. doi:10.1007/s11192-013-0978-8

Zhao, Z., Rollins, J., Bai, L., & Rosen, G. (2017, Oct. 2017). *Incremental Author Name Disambiguation for Scientific Citation Data.* Paper presented at the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Zhu, J., Wu, X., Lin, X., Huang, C., Fung, G. P. C., & Tang, Y. (2018). A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering. *Scientometrics, 114*(3), 781-794. doi:10.1007/s11192-017-2611-8

Title: Ethnicity-Based Name Partitioning for Author Name Disambiguation Using Supervised Machine Learning

Authors: Jinseok Kim, Jenna Kim, and Jason Owen-Smith

1. Jinseok Kim

Institute for Research on Innovation & Science, Survey Research Center, Institute for Social Research
University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104
jinseokk@umich.edu
ORCID id: 0000-0001-6481-2065

2. Jenna Kim

School of Information Sciences
University of Illinois at Urbana-Champaign
501 E. Daniel Street, Champaign, IL U.S.A. 61820
jkim682@illinois.edu

ORCID id: 0000-0001-7438-448X

3. Jason Owen-Smith

Department of Sociology, Institute for Social Research, University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104-2910
734-936-0463
jdos@umich.edu
ORCID id: 0000-0001-8007-1121

[Corresponding Author Information]

Jinseok Kim

Institute for Research on Innovation & Science, Survey Research Center, Institute for Social Research
University of Michigan
330 Packard Street, Ann Arbor, MI U.S.A. 48104-2910
734-763-4994
jinseokk@umich.edu

[Financial Support Information]

This paper was supported by grants from the National Science Foundation (#1917663, #1760609, and #1561687).

Abstract

In several author name disambiguation studies, some ethnic name groups such as East Asian names are reported to be more difficult to disambiguate than others. This implies that disambiguation approaches might be improved if ethnic name groups are distinguished before disambiguation. We explore the potential of ethnic name partitioning by comparing performance of four machine learning algorithms trained and tested on the entire data or specifically on individual name groups. Results show that ethnicity-based name partitioning can substantially improve disambiguation performance because the individual models are better suited for their respective name group. The improvements occur across all ethnic name groups with different magnitudes. Performance gains in predicting matched name pairs outweigh losses in predicting nonmatched pairs. Feature (e.g., coauthor name) similarities of name pairs vary across ethnic name groups. Such differences may enable the development of ethnicity specific feature weights to improve prediction for specific ethic name categories. These findings are observed for three labeled data with a natural distribution of problem sizes as well as one in which all ethnic name groups are controlled for the same sizes of ambiguous names. This study is expected to motive scholars to group author names based on ethnicity prior to disambiguation.

Keywords: supervised machine learning; author name disambiguation; feature engineering, name ethnicity; data partition

Introduction and Background

A big challenge in managing digital libraries is that author names in bibliographic data are ambiguous because many authors have the same names (homonyms) or variant names are recorded for the same authors (synonyms). One study estimates that about two-thirds of author names in PubMed, the largest biomedicine digital library, are vulnerable to either or both of these two ambiguity types (Torvik & Smalheiser, 2009). Research findings obtained by mining bibliographic data can be distorted by merged and/or split author identities due to incorrect disambiguation (Fegley & Torvik, 2013; J. Kim & Diesner, 2015, 2016; Schulz, 2016). In addition, digital library users query author names most frequently (Islamaj Dogan, Murray, Névéol, & Lu, 2009). This means that the users will receive inaccurate information about research production, citation, and collaboration for authors if author name ambiguity is not properly resolved (Harzing, 2015; Strotmann & Zhao, 2012).

To address the challenge, researchers have proposed a variety of author name disambiguation (AND, hereafter) methods. Some scholars have used heuristics such as string-based matching (e.g., names that have the same full surname and forename initials are assumed to represent the same author), which is the most widely used approach in bibliometrics (Milojević, 2013). Others have developed rule-based programming and supervised/unsupervised machine learning techniques, as systemically reviewed in several papers (Ferreira, Gonçalves, & Laender, 2012; Hussain & Asghar, 2017; Sanyal, Bhowmick, & Das, 2019; Smalheiser & Torvik, 2009). In industry, several bibliographic data providers such as DBLP, Scopus, and Web of Science have disambiguated author names to improve their service quality (Kawashima & Tomizawa, 2015; J. Kim, 2018; Ley, 2009; Zhao, Rollins, Bai, & Rosen, 2017), while others still rely on the name string matching to output author-related search results.

Despite the differences in methods and datasets, a few AND studies have observed that some ethnic name groups (ENG, hereafter) (e.g., Chinese names) are more difficult to disambiguate than others (Deville et

al., 2014; J. Kim & Diesner, 2016; Strotmann & Zhao, 2012; Torvik & Smalheiser, 2009; J. Wu & Ding, 2013). This implies that author names may be better disambiguated if their associated ethnicities are considered as inputs in disambiguation models. But this possibility has been little explored. First, the observations made in several studies that certain ethnic names are harder to disambiguate are based on post-hoc evaluations of AND results. In other words, many of those studies did not integrate ethnic name partitions during machine learning. A very small number of studies have divided names into subgroups in their disambiguation model building (Chin et al., 2013; Louppe, Al-Natsheh, Susik, & Maguire, 2016) and evaluation process (Lerchenmueller & Sorenson, 2016). But their ethnic name categories are limited in number (e.g., dichotomy of Chinese vs non-Chinese; Caucasian, Asian, and Hispanic) or mixed up with racial distinctions based on the U.S. Social Security information (e.g., White, Black, Hispanic, Asian, etc.). Such racial classifications can be inappropriate for bibliographic data in which author names come from diverse regions around the world. In addition, those studies have typically used a single labeled data source, which makes it hard to expand and generalize their findings to other AND scenarios.

This study aims to empirically evaluate the effect ethnic name partitioning has on AND. In this study, author name disambiguation is a task to assign either ‘match’ or ‘nonmatch’ label to a pair of author name instances. For this, specifically, name instances are grouped into a block that share the same first forename initial and full surname and pairwisely compared with the block for their similarities over a set of features (e.g., coauthor name) to produce similarity scores. Machine learning algorithms combine the scores to learn weights of each feature to decide if a given pair of instances to refer to the same author or not. Although our work is motivated by the studies reviewed above and follows their common data pre-processing, blocking and machine learning steps, this paper differs from them in three important ways. First, this study evaluates AND performance by four different machine learning algorithms applied to four different labeled datasets before and after inclusion of a standard ethnic name group partition. Here, a name instance is assigned to an ethnic name group based on a name ethnicity classification system, *Ethnea*. Second, unlike traditional labeled data in which a specific ENG (i.e., Chinese) dominates, this study disambiguates new labeled data in which all ENGs are controlled to have the same numbers of instances, to demonstrate that performance changes induced by ethnic name partitioning may not be solely due to the well-known relationship between the number of cases and their ambiguity (more names, more ambiguity). Third, this study shows that different combinations of features (e.g., coauthor name and title words) appear to be related to AND performance for different ENGs suggesting future directions to further improve AND performance with ambiguous ethnic group names. The findings of this study can provide practical insights to researchers and practitioners who handle authority control in digital libraries. In the following sections, details on labeled data and setups for machine learning are described.

Method

Labeled Data and Pre-Processing

To measure the effect of ethnic name partition on machine learning for AND, this study disambiguates names in four labeled datasets – KISTI, AMINER, GESIS, and UM-IRIS. The first three datasets have been used in many AND studies to train and test machine learning algorithms (Cota, Ferreira, Nascimento, Gonçalves, & Laender, 2010; Ferreira, Veloso, Gonçalves, & Laender, 2014; Hussain & Asghar, 2018; J. Kim & Kim, 2018, In print; Momeni & Mayr, 2016; Alan Filipe Santana, Gonçalves, Laender, & Ferreira, 2017; Shin, Kim, Choi, & Kim, 2014; H. Wu, Li, Pei, & He, 2014; Zhu et al., 2018).

The last one is added to investigate how the ethnic name partition affects AND under the condition in which all ENGs are constrained to have the same numberss of ambigous name instances[1].

KISTI: Scientists at the Korea Institute of Science & Technology Information (KISTI) and Kyungsung University in Korea constructed this labeled dataset. It is made up of 41,673 author name instances that belong to 6,921 unique authors (Kang, Kim, Lee, Jung, & You, 2011)[2].

AMINER: Researchers in China and U.S. collaborated to create this labeled data to build and evaluate AND models for a computer science digital library, AMiner (Tang et al., 2008; X. Wang, Tang, Cheng, & Yu, 2011)[3]. It consists of 7,528 author name instances that refer to 1,546 unique authors.

GESIS: Scholars at the Leibniz Institute for the Social Sciences (GESIS) in Germany produced this labeled data. It contains author name instances of 5,408 unique authors (Momeni & Mayr, 2016) [4]. This study reuses the 'Evaluation Set' (29,965 author name instances of 2,580 unique authors) but with a few enhancements (J. Kim & Kim, 2020). Each author name instance is converted into the 'surname, forename' format and, through linking GESIS to its base DBLP data, is associated with the title of the paper in which it appears and the name of the conference or journal where the paper is published.

UM-IRIS: This dataset was generated by the researchers at the University of Michigan Institute for Research on Innovation & Science (UM-IRIS) and the University of Illinois through matching selected name instances in publication records to an authority database, ORCID (Kim & Owen-Smith, in print). First, author full names (e.g., 'Brown, Michael') that appear 50 times or more in MEDLINE-indexed publications published between 2000 and 2019 were listed[5]. Then, all instances of each selected full name (e.g., 158 instances of 'Brown, Michael' in MEDLINE) and their associated publication metadata were compared to 6 million researcher profiles in ORCID[6]. If an instance had a single match in the publication list of an ORCID researcher profile (matching on full name, paper title, and publication venue), the matched researcher's ORCID id was assigned as an author label to the instance. Next, among the ORCID id-linked instances, those whose full names are associated with 5 or more ORCID ids (e.g., 6 unique ORCID researchers share the name 'Brown, Michael' which appear 158 times in MEDLINE) were randomly selected to produce 1,000 name instances for each of six ENGs. The resulting data contain 6,000 instances of 822 authors.

Four features – author name, coauthor name(s), paper title, and publication venue - are used as machine learning features because they have been widely used in algorithmic AND studies (Schulz, 2016; Song, Kim, & Kim, 2015) and are commonly available in the four labeled datasets. The string of each feature is stripped of non-alphabetical characters, converted into ASCII format, and lowercased. For title words, common English words like 'the' and 'to' are removed (i.e., stop-word listed) using the dictionary in Stanford NLP[7] and stemmed (e.g., 'solution' → 'solut') using the Porter's algorithm[8]. Name instances in KISTI are converted into the full surname and first forename initial format ('Wang, Wei' → 'Wang, W') to make them more ambiguous (see J. Kim & Kim, 2020).

[1] This new labeled dataset was created following the idea of a reviewer who suggested that the impact of ethnic name partition on AND may be confounded by the size differences of ambiguous names.

[2] http://www.lbd.dcc.ufmg.br/lbd/collections/disambiguation/DBLP.tar.gz/at_download/file

[3] http://arnetminer.org/lab-datasets/disambiguation/rich-author-disambiguation-data.zip

[4] http://dx.doi.org/10.7802/1234

[5] https://www.nlm.nih.gov/bsd/medline.html

[6] https://orcid.org/

[7] https://github.com/stanfordnlp/CoreNLP/blob/master/data/edu/stanford/nlp/patterns/surface/stopwords.txt

[8] https://tartarus.org/martin/PorterStemmer/

Ethnic Name Group (ENG) Tagging

This study assigns an ENG tag to a name instance in each labeled dataset using the author name ethnicity classification database, *Ethnea*, developed by Torvik and Agarwal (2016)[9]. *Ethnea* is a collection of more than 9 million author name instances that are tagged one of 26 ENG classes based on the name's association with national-level geo-locations[10]. For example, "Wang, Wei" is classified as 'Chinese' as it is most frequently associated with organizations in China. However, *Ethnea* makes no distinctions based on any anthropological, cultural, or linguistic characteristics of authors. Instead it relies entirely on observations of names and geo-locations of their frequently associated institutions so an author named 'Wang, Wei' who was born in the U.S. and has never visited China would still be assigned a 'Chinese name' tag. We link all four labelled datasets to *Ethnea* and, if a matched name is found, assign that name's ENG tag to all its observed instances. If a queried name does not have a match in *Ethnea*, we search again using only the surname and assign the modal *Ethnea* ENG tag associated with it to all its instances. Table 1 summarizes the frequencies and ratios of ENG tags assigned by *Ethnea* to author name instances in each labeled dataset.

Table 1 shows the list of ENGs in each labeled data. Small-sized ENGs are excluded from analysis because most name instances in those ENGs tend to belong to a single author while a few instances referring to other author(s). When randomly split into training and test subsets for machine learning, these instances do not produce negative pairs at all.

Table 1: Summary of Ethnic Name Group (ENG) Frequencies in Labeled Data

Labeled Data (No. of Instances)							
KISTI (41,605)		**AMINER (7,528)**		**GESIS (29,965)**		**UM-IRIS (6,000)**	
ENG	**Ratio (%)**	**ENG**	**Ratio (%)**	**ENG**	**Ratio (%)**	**ENG**	**Ratio (%)**
CHINESE	50.1	CHINESE	64.2	CHINESE	56.2	CHINESE	16.67
ENGLISH	15.6	ENGLISH	17.0	GERMAN	14.8	ENGLISH	16.67
INDIAN	9.5	INDIAN	6.3	ENGLISH	7.0	GERMAN	16.67
KOREAN	8.0	GERMAN	5.6	INDIAN	4.3	HISPANIC	16.67
GERMAN	3.3	HISPANIC	3.1	HISPANIC	3.6	INDIAN	16.67
ISRAELI	2.2	**Sum**	**96.2**	KOREAN	3.5	KOREAN	16.67
ITALIAN	2.0	Excluded ENGs (3.8%): JAPANESE, NORDIC, KOREAN, ARAB		JAPANESE	2.8	**Sum**	**100.00**
HISPANIC	1.7			ITALIAN	1.7	Excluded ENGs (0%): None	
JAPANESE	1.1			ARAB	1.6		
DUTCH	1.0			FRENCH	1.3		
ARAB	0.9			**Sum**	**96.8**		
FRENCH	0.9			Excluded ENGs (3.2%): NORDIC, HUNGARIAN, DUTCH, VIETNAMESE, SLAV, ROMANIAN, GREEK, ISRAELI, TURKISH, INDONESIAN			
Sum	**96.3**						
Excluded ENGs (3.7%) : NORDIC, SLAV, GREEK, ROMANIAN, VIETNAMESE, NULL, AFRICAN, TURKISH, HUNGARIAN,							

[9] https://databank.illinois.edu/datasets/IDB-9087546

[10] 26 ethnicities include: African, Arab, Baltic, Caribbean, Chinese, Dutch, English, French, German, Greek, Hispanic, Hungarian, Indian, Indonesian, Israeli, Italian, Japanese, Korean, Mongolian, Nordic, Polynesian, Romanian, Slav, Thai, Turkish, and Vietnamese. In *Ethnea*, some name instances are assigned two ethnicities (e.g., "Jane Kim" → Korean-English) if the surname and forename of an author name are associated frequently with different ethnicities.

Chinese names represent the majority of ENGs in three labeled data. This is because these labeled data were created from computer science papers where Chinese researchers are particularly large contributors. In addition, as the three datasets were designed to collate challenging names to disambiguate, Chinese names that tend to be more ambiguous than other ENGs were over-sampled (Müller, Reitz, & Roy, 2017). In contrast, 6,000 instances in UM-IRIS are evenly distributed over six ENGs. For validation and reuse, these labeled data with ENG tags are publicly available[11]. Note that the original KISTI contains 41,673 name instances, whereas the ENG-tagged KISTI has 41,605 instances. Such discrepancy occurs because this paper uses the revised version of KISTI that corrects record errors and duplicates in the original data (J. Kim, 2018).

Machine Learning Process

Machine learning methods for AND can be divided into two groups: author assignment and author grouping (Ferreira et al., 2012). While the former aims to assign an author name instance to one of pre-disambiguated author name clusters, the latter aims to group all and only instances that belong to the same authors. This study takes the latter approach in evaluating the effect of ENG on AND. Specifically, author name instances in each labeled dataset are pairwise compared to assess whether a given instance pair of plausibly represents the same author (a match) or not (a nonmatch). Although some scholars take a further step to cluster pairwise comparisons (e.g., J. Kim & Kim, 2018; Levin, Krawczyk, Bethard, & Jurafsky, 2012; Louppe et al., 2016; Alan Filipe Santana et al., 2017), this study only evaluates disambiguation performance at a pair level (i.e., classification), following the practice of previous AND studies (e.g., Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004; Song et al., 2015; Treeratpituk & Giles, 2009; Vishnyakova, Rodriguez-Esteban, Ozol, & Rinaldi, 2016).

As the first machine learning step, author name instances in each labeled dataset are randomly divided into training (50%) and test (50%) subsets. Then, instances in each subset are put into blocks in which all member instances share the same full surname and first forename initial (e.g., 'Wang, W'). Only instances in the same block are compared for disambiguation. This blocking is typical in AND studies because it reduces computational complexity with only slight performance degradation (K. Kim, Sefid, & Giles, 2017; Torvik & Smalheiser, 2009). Next, instance pairs in the same block are compared to establish their similarity over four other data features: author name, coauthor name(s), paper title, and publication venue. To quantify how much a pair is similar over a feature, this study calculates the cosine similarity of Term (*n*-gram) Frequency for each feature (Han, Zha, & Giles, 2005; J. Kim & Kim, In print; Levin et al., 2012; Louppe et al., 2016; A. F. Santana, Gonçalves, Laender, & Ferreira, 2015; Treeratpituk & Giles, 2009). Specifically, the string of a feature is converted into an array of 2~4-grams (e.g., author name 'Wang, Wei' → 'wa|an|ng|gw|we|ei|wan|ang|ngw|gwe|wei|wang|angw|ngwe|gwei'). After the conversion, two *n*-gram arrays of an instance pair are compared to produce a cosine similarity score for the feature.

Besides the four basic features, ENGs are used as a feature set for ENG-aware disambiguation. For this, especially, an instance pair's ENG is encoded into a binary value (i.e., one-hot encoding) for a pre-defined set of ENGs[12]. For example, in AMINER, a pair of name instances ('Wang, Wei' and 'Wang, W.') is assigned either 'Yes' or 'No' for each of five ethnicities – Chinese ('Yes'), English ('No'), Indian

[11] Download link TBA

[12] As only instances in the same block in which they share at least the same full name and first forename initial are compared, all the pairs in the block have the same ethnicity tag.

(‘No’), German (‘No’), and Hispanic (‘No’) – as shown in Table 1. Table 2 shows examples of the cosine similarity scores calculated over four features and ENG encoding results for instance pairs.

Table 2: A Mock-Up Example of Cosine Similarity Scores for Instance Pairs over Four Features and ENG Encoding

Pairs	Feature								Label
	Author Name	Coauthor Name	Paper Title	Pub. Venue	ENG 1	ENG 2	ENG 3	…	
Pair 1	0.97	0.89	0.67	0.12	Yes	No	No	…	Match
Pair 2	1.00	0.24	0.46	0.00	No	Yes	No	…	Nonmatch
Pair 3	0.65	0.07	0.00	0.80	No	No	Yes	…	Nonmatch
Pair 4	0.58	0.08	0.00	0.00	Yes	No	No	…	Nonmatch

We focus on four algorithms – Gradient Boosting, Logistic Regression, Naïve Bayes, and Random Forest – for supervised machine learning that have been widely used as baselines or best performing methods in AND studies (e.g., Han et al., 2004; J. Kim & Kim, In print; K. Kim, Sefid, Weinberg, & Giles, 2018; Louppe et al., 2016; Song et al., 2015; Torvik & Smalheiser, 2009; Treeratpituk & Giles, 2009; Vishnyakova et al., 2016; J. Wang et al., 2012). In the first scenario, they are trained on the list of similarity scores and labels, as shown in Table 2 to learn relative weights for features and an absolute weight or threshold for instance pairs to be disambiguated without considering ENGs (→ ENG-ignorant learning). In the second scenario, the same algorithms are trained on the list of similarity scores, ENGs, and labels (→ ENG-aware disambiguation). Here, the ethnic name partition adds more features (dimensions) to each instance pair’s feature set, allowing algorithms to combine the similarities of the expanded features. The machine learning procedure is implemented using the python Scikit-learn package. For Gradient Boosting, 500 estimators are used with max depth=9 and learning rate = 0.125. For Logistic Regression, L2 Regularization with class weight = 1 is used. Gaussian Naïve Bayes with maximum likelihood estimator is used for Naïve Bayes. For Random Forest, 500 trees are used after a grid search.

Trained algorithmic models are applied to the instance pairs in test subsets in which the cosine similarity is calculated for the four basic features and, in the second scenario, ethnicities are encoded in the same fashion but explicitly include ENG information. As in Table 2, an algorithmic model receives a set of feature similarity scores and, if ENG-aware disambiguation is conducted, a list of encoded ENGs for an instance pair to output a binary classification decision (match or nonmatch). Once trained, each algorithm produces a single score that predicts the probability of an instance pair being negative (nonmatch). If the predicted probability is above a certain threshold (> 0.5), the pair is decided to be a nonmatch, whereas if below the threshold, a match.

Performance Evaluation

We evaluate each algorithm’s classification results on reserved test subsets of each labeled dataset by calculating precision and recall for positive (P; match) and negative (N; nonmatch) pairs respectively. In addition, we calculate the F1 score as a harmonic mean of precision and recall.

Specifically, precision for positive pairs (Prec-Pos) measures how many predicted match pairs are correct ones (true positives; TP) over the total number of predicted match pairs that may contain correct match pairs (true positives; TP) and incorrect match pairs (false positives; FP). In contrast, recall for positive pairs (Rec-Pos) measures the ratio of correct match pairs (true positives; TP) over the total number of true match pairs that may be predicted correctly as match pairs (true positives; TP) or incorrectly as nonmatch pairs (false negatives; FN).

$$Precision\ Positive\ (PrecPos) = \frac{Number\ of\ Correctly\ Predicted\ Match}{Number\ of\ Predicted\ Match} = \frac{TP}{(TP + FP)} \qquad (1)$$

$$Recall\ Positive\ (RecPos) = \frac{Number\ of\ Correctly\ Predicted\ Match}{Number\ of\ True\ Match} = \frac{TP}{(TP + FN)} \qquad (2)$$

$$F1\ Positive = \frac{2 \times PrecPos \times RecPos}{PrecPos\ +\ RecPos} \qquad (3)$$

Likewise, precision for negative pairs (Prec-Neg) measures how many predicted nonmatch pairs are correct ones (true negatives; TN) over the total number of predicted nonmatch pairs that may contain correct nonmatch pairs (true negatives; TN) and incorrect nonmatch pairs (false negatives; FN). In contrast, recall for negative pairs (Rec-Neg) measures the ratio of correct nonmatch pairs (true negatives; TN) over the total number of true nonmatch pairs that may be predicted correctly as nonmatch pairs (true negatives; TN) or incorrectly as match pairs (false positives; FP).

$$Precision\ Negative\ (PrecNeg) = \frac{Number\ of\ Correctly\ Predicted\ Nonmatch}{Number\ of\ Predicted\ Nonmatch} = \frac{TN}{(TN + FN)} \qquad (4)$$

$$Recall\ Negative\ (RecNeg) = \frac{Number\ of\ Correctly\ Predicted\ Nonmatch}{Number\ of\ True\ Nonmatch} = \frac{TN}{(TN + FP)} \qquad (5)$$

$$F1\ Negative = \frac{2 \times PrecNeg \times RecNeg}{PrecNeg\ +\ RecNeg} \qquad (6)$$

The metrics are calculated on the entire set of test results for each labeled dataset. We separately calculate performance measures for difference ENGs rather than averaging them across multiple ethnicity groups.

Results

Cross-Data Performance Evaluation

Figure 1 shows disambiguation results on KISTI, reporting precision and recall *before* and *after* ENG-aware disambiguation by four algorithms – Gradient Boosting (GB), Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF). Figure 1a shows that when ENGs are included as features, the algorithms tend to produce better precision in the prediction of positive (match) pairs than when they are not considered. This is shown by black bars ('After') being higher than stripped bars ('Before') in Figure 1a. This observation indicates that ethnic name partitioning helps algorithms increase the ratio of TP among predicted positive pairs (= TP + FP). This can be confirmed by checking the numbers of true and false positive pairs in Table 3. For example, when trained only on the four basic (non-ENG) features, LR predicts that 76,201 (= TP + FP = 55,998 + 20,203) pairs refer to the same authors (match) and 73.49% of the predictions are right (= TP/(TP + FP)). After trained on the same but ENG-tagged data, however, it predicts 170,432 pairs to be match sets, increasing its prediction accuracy this time to 77.08%.

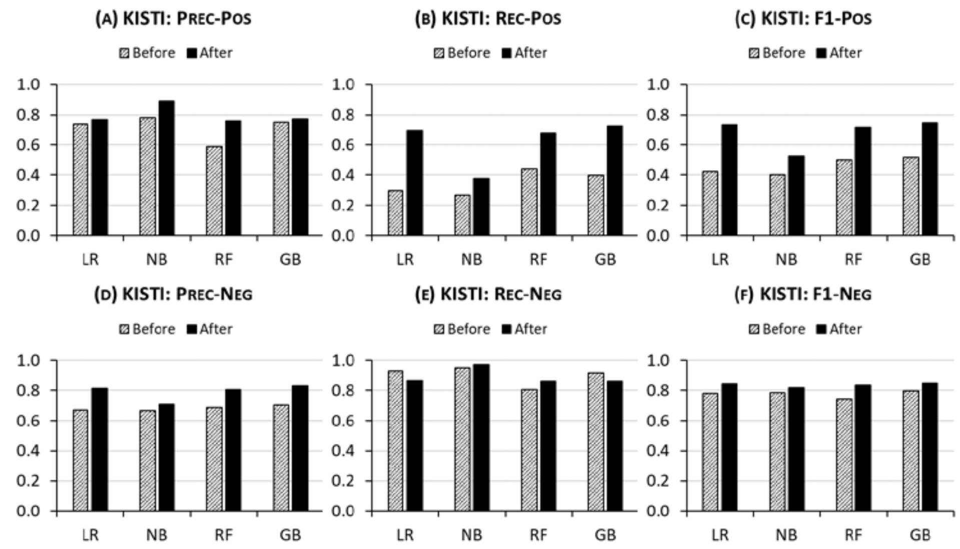


Figure 1: Disambiguation Performances 'Before' Versus 'After' ENG-Aware Disambiguation on KISTI

Table 3: Numbers of Correctly or Incorrectly Predicted Pairs for Positive and Negative Pairs by Four Algorithms on KISTI Test Data

Algorithm	ENGs Considered	No. of Pairs (P: Positive) (N: Negative)	TP	FN	FP	TN
LR	Before	483,029 (P: 189,375) (N: 293,654)	55,998	133,377	20,203	273,451
	After		131,364	58,011	39,068	254,586
NB	Before		50,782	138,593	14,267	279,387
	After		71,020	118,355	8,775	284,879
RF	Before		82,727	106,648	57,532	236,122
	After		128,105	61,270	40,479	253,175
GB	Before		75,195	114,180	24,725	268,929
	After		136,859	52,516	39,888	253,766

Ethnic name partitioning also reduces the number of falsely predicted nonmatch cases (FN), increasing recall in Figure 1b. Performance gains by ENG-aware disambiguation are more pronounced for recall than for precision, as evidenced by larger differences between 'Before' and 'After' bars for recall (Figure 1b) than those for precision (Figure 1a). In other words, ENG aware disambiguation across four common algorithms appears to reduce false negative predictions more than true positive predictions, potentially providing better performance for applications (such as network analysis) that are particularly sensitive to biases due to erroneous "lumping" of name instances that actually refer to different individuals. The improvements in precision and recall together increase the F1 scores by ENG-aware disambiguation (Figure 1c).

ENG-aware disambiguation also does a better job of accurately predicting non-match (negative pair) cases. The 'After' bars are taller than those of 'Before' in Figure 1d. Unlike the positive pair prediction in which ENG-aware disambiguation works in favor of both precision and recall by all algorithms, however, the performance gains in precision for negative pairs come with slightly decreased recall by GB and LR in Figure 1e. This means that while disambiguation models by GB and LR trained on ENG-added features are good at increasing the numbers of true nonmatch pairs among predicted nonmatch pairs (= TN +FN), they incorrectly predict that true nonmatch pairs match (FP predictions) more frequently than when they are trained on the four basic features alone. Reduced recall for negative pair prediction is, however, offset by increased precision, leading to the F1 scores by ENG-aware disambiguation being better than those by ENG-blind one in Figure 1f. Meanwhile, NB and RF still obtain improvements in both precision and recall as well as F1.

Algorithmic performances are also enhanced by ENG-aware disambiguation on AMINER, GESIS, and UM-IRIS. Figure 2 ~ 4 report that the algorithms trained on ENG-tagged data perform better than those trained only on the basic features across almost all metrics for both positive and negative pairs. NB models prove the exception, producing worse results in recall for positive pairs and in precision for negative pairs after ethnic name partition. However, this degraded performance is offset by increased precision for positive pairs and increased recall for negative pairs, respectively, so the overall performance metric (F1), which equally weights precision and recall, indicates an overall improvement due to the inclusion of ENG features.

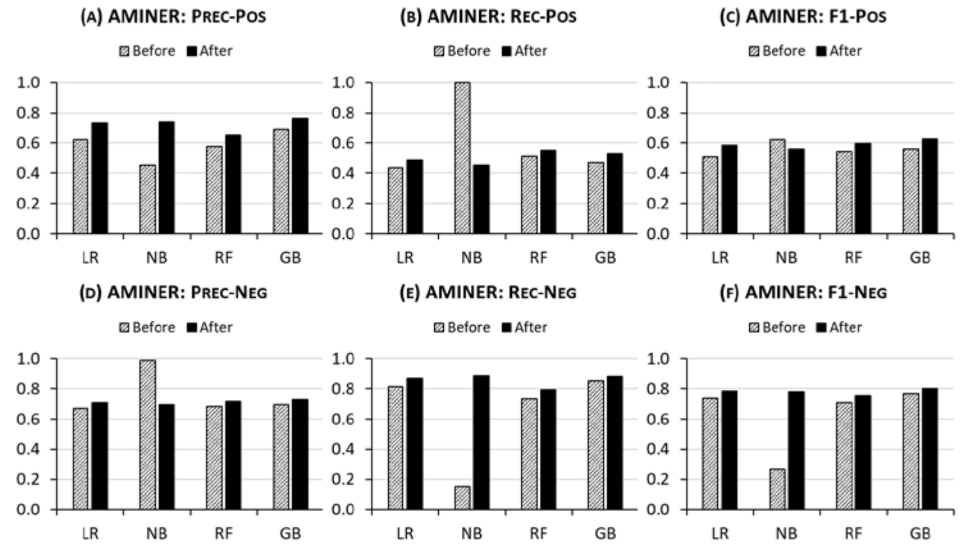


Figure 2: Disambiguation Performances 'Before' Versus 'After' ENG-Aware Disambiguation on AMINER

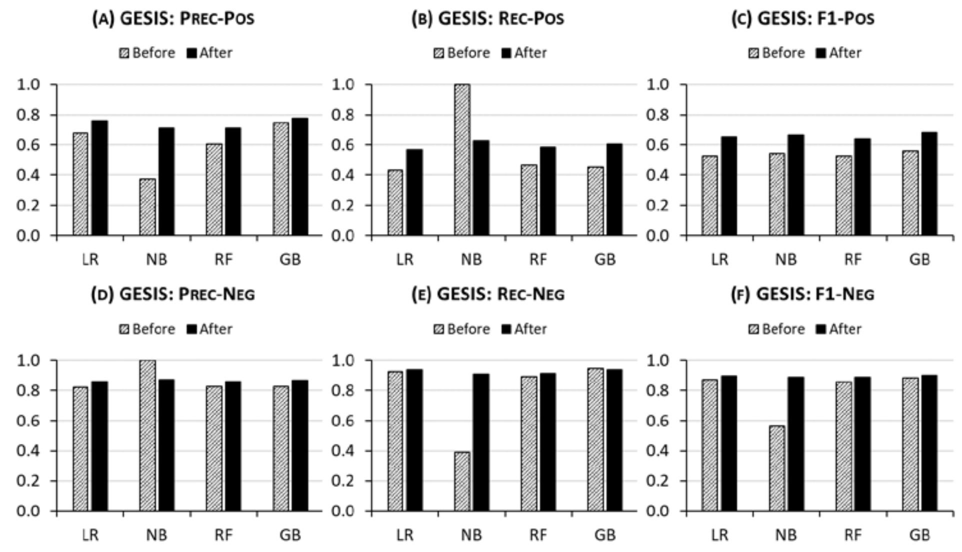


Figure 3: Disambiguation Performances 'Before' Versus 'After' ENG-Aware Disambiguation on GESIS

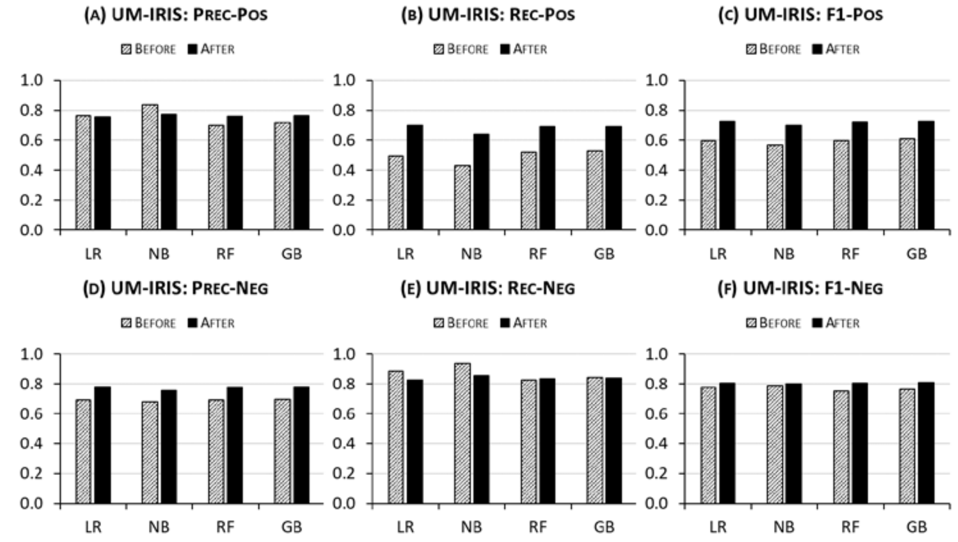


Figure 4: Disambiguation Performances 'Before' Versus 'After' ENG-Aware Disambiguation on UM-IRIS

Performance Evaluation per ENG

ENG-aware disambiguation produces substantial improvements in both precision and recall for predicting match and nonmatch instance pairs in different labeled datasets. But are those improvements uniform across different ENGs? If not, a more nuanced approach to model evaluation may be necessary. To answer this question, we compare performance changes due to ENG-aware disambiguation within ENG groups. For this, precision, recall, and F1 scores for positive and negative pairs predicted by four algorithms are calculated separately for instance pairs that belong to the same ENG in each of four labeled data: 4 algorithms × 4 data = 16 evaluations. Presenting all the results at the same time would consume too much space in this paper. So, we present random forest (RF) predictions on the GESIS dataset as an illustration for the purposes of this discussion. Reports of other algorithms and data are presented in a supplementary document attached to this paper.

Figure 5 shows the by ENG performance metrics for the RF algorithm trained on GESIS with and without ENG-aware disambiguation. The ENG-aware disambiguation leads to better precision (positive pairs; Figure 5A) and recall (negative pairs; Figure 5E) for Chinese names but worse precision (positive pairs) and recall (negative pairs) for other ethnicities. In contrast, name disambiguation for Chinese names results in lower recall (positive pairs; Figure 4B) and precision (negative pairs; Figure 4D) than those for other ENGs. Similar patterns are observed for other algorithms tested on GESIS (see Figure S5~S8 in Supplementary Material). This suggests that the effect of ENG-aware disambiguation occurs in different ways for different ENGs. Thus, its application can be beneficial in some instances but detrimental in others. Variations in the effects of ENG-aware disambiguation on precision and recall for positive and negative pair prediction across ethnicity groups suggest that care must be taken to design disambiguation strategies that fit particular analytic or empirical needs.

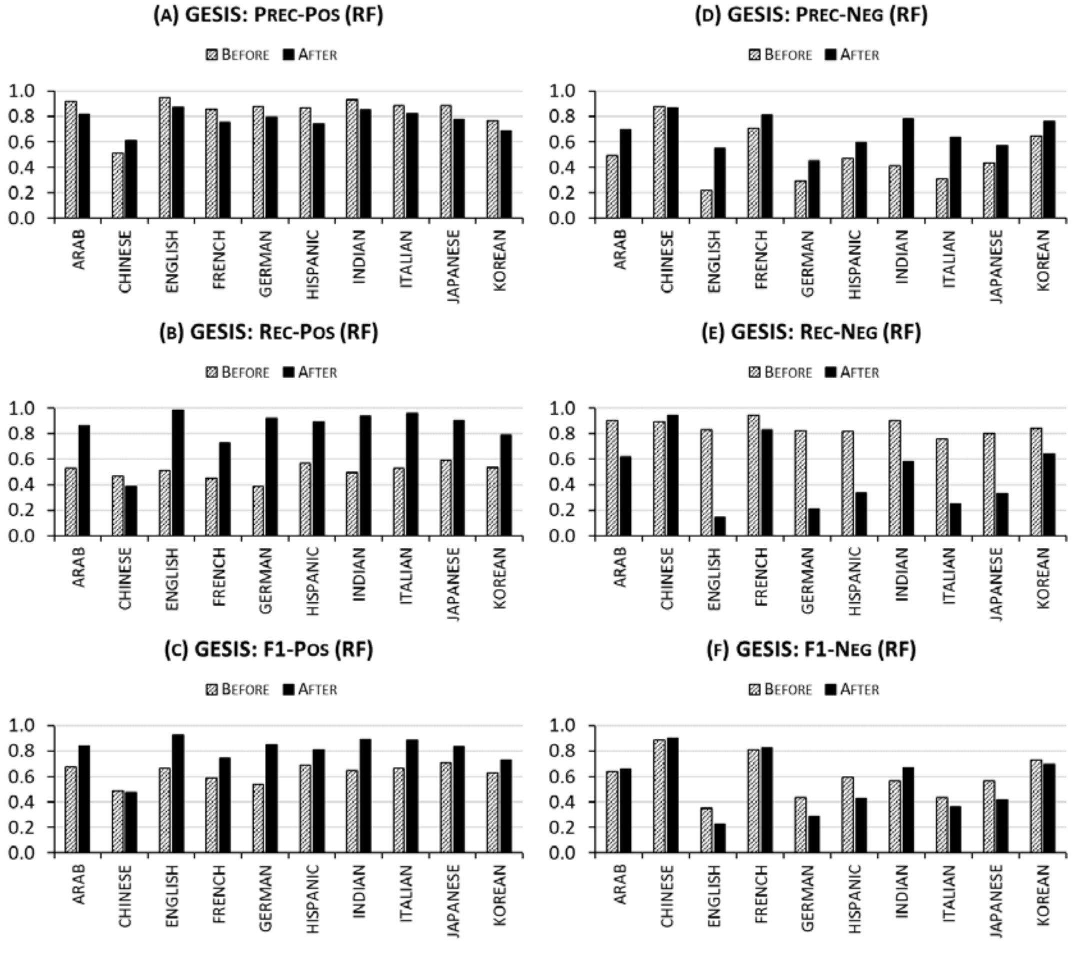


Figure 5: Disambiguation Performances per ENG 'Before' Versus 'After' ENG-Aware Disambiguation by Random Forest on GESIS

These observations can be explained as follows. ENGs have different distributions of similarity scores over the four basic (non-ethnicity) features we use. Figure 6 presents the feature similarity score distributions per ENG for positive (left) and negative (right) pairs in the GESIS test data. Training and test subsets show similar distributions in each labeled dataset. For visual simplicity, a score is rounded up into nearest bins with intervals of 0.1 on *x*-axis and the ratios of the numbers of scores in the same bin over all scores are plotted on *y*-axis. A solid red line represents the distribution of all instance pairs regardless of ENG.

In Figure 6, each ENG has different distributions of, for example, 'COAUTHOR' similarity scores for both positive and negative pairs (Figure 6C and 6D). So, the four algorithms come to use different 'coauthor' similarity score distributions in ENG-aware disambiguation. Such heterogeneous distributions also occur for other features but with different variations of differences. For example, 'VENUE' distributions in Figure 6G and 6H differ less across ENGs than do 'COAUTHOR' distributions. Because ENG-aware disambiguation allows training and testing on different feature similarity score distributions for each ENG, the algorithms combine features using different weightings for each ethnicity, producing different predictions for name pairs with the same feature similarity scores but different ENG tags. In other words, this method takes into account the likelihood that researchers in different ENGs organize their scientific work differently, favoring distinct co-authorship and publication venue patterns. This also occurs in disambiguation of other labeled data, whose feature similarity score distributions are reported in Figure S17 ~ S20 in Supplementary Material.

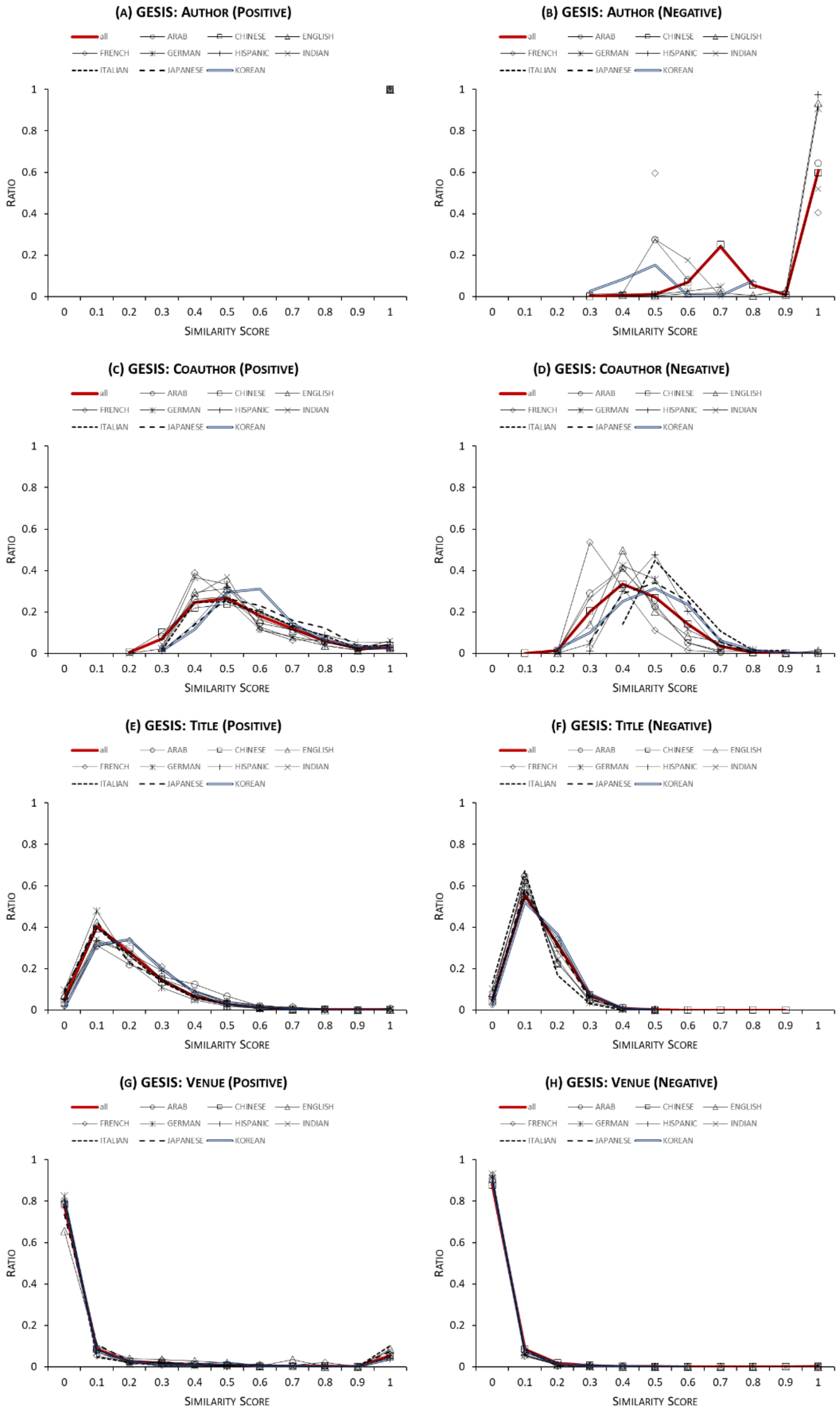


Figure 6: Feature Similarity Score Distributions per ENG for Positive and Negative Pairs in GESIS Test Data

Figure 5 also shows that some ENGs manifest substantial improvements in recall for positive pairs (Figure 5B) but degraded recall for negative pairs (Figure 5E). This might be explained in two ways. In our 'before' (ENG-unaware) case, algorithms combine features to produce per-feature weights for positive pairs based on feature similarity scores aggregated across multiple ENGs that can have very different feature distributions. Such aggregated distributions cannot effectively capture the single match patterns specific to each ENG, which seem to lead models to falsely predict positive pairs as negative ones (FN), reducing the recall for positive pairs. Conversely, increased recall for positive pairs after ENG-aware disambiguation means that the algorithms trained and tested on ENG-tagged data successfully produce per-feature weights optimized to each ENG, thus making better predictions that push up the recall scores for many ENGs.

Second, decreased negative pair recall after ENG-aware disambiguation means that the algorithms trained and tested on ENG-tagged data fail to produce proper per-feature weights for accurately predicting nonmatch for known negative pairs. When the algorithms are trained only on the four basic (non-ethnicity) features, they do a better job of predicting nonmatch pairs based on aggregated feature similarity distributions that are invariant across particular ENGs. In other words, feature distributions aggregated across ENGs appear to be more effective for predicting negative case pairs while ENG-aware disambiguation techniques more accurately capture positive pairs.

These observations imply that disambiguation models for positive pair prediction would be improved by ENG-aware procedures, while nonmatch patterns for negative pair prediction can aggregate across ENGs (J. Kim & Kim, 2018). Table 4 shows that in the GESIS training data, each ENG has different sizes of positive and negative (pairwise) pairs. CHINESE name instances produce the largest numbers of positive (≈ 146K) and negative pairs (≈ 551K), while ITALIAN name instances generate around a few thousand positive and a few hundred negative pairs. In other training data, CHINESE pairs constitute substantially large proportions (KISTI: 71.28 % and AMINER: 91.26 %) or over one-third (UM-IRIS: 37.54 %) of all negative pairs, while other ENG pairs make up small or less-than-expected (approximately 17% per ENG in UM-IRIS) proportions. In contrast, the numbers of positive pairs are less concentrated (GESIS, KISTI, and AMINER) or more evenly distributed (UM-IRIS) for positive pairs than those for negative pairs.

Table 4: Distributions of Positive and Negative Name Instance Pairs per ENG in GESIS Training Data

ENG	Positive Pairs	Ratios	Negative Pairs	Ratios
ARAB	1,559	0.66	1,203	0.21
CHINESE	145,969	62.22	551,150	93.96
ENGLISH	16,020	6.83	2,147	0.37
FRENCH	1,948	0.83	3,118	0.53
GERMAN	37,373	15.93	13,092	2.23
HISPANIC	5,887	2.51	1,907	0.33
INDIAN	6,708	2.86	2,654	0.45
ITALIAN	3,489	1.49	686	0.12
JAPANESE	8,853	3.77	2,608	0.44
KOREAN	6,792	2.90	7,994	1.36
Total	234,598	100	586,559	100

Table 5: Distributions of Positive and Negative Name Instance Pairs per ENG in KISTI Training Data

ENG	Positive Pairs	Ratios	Negative Pairs	Ratios
ARAB	6,015	3.21	37	0.01
CHINESE	53,519	28.55	209,807	71.28
DUTCH	5,448	2.91	84	0.03
ENGLISH	45,054	24.04	25,137	8.54
FRENCH	2,686	1.43	625	0.21
GERMAN	11,065	5.90	4,114	1.40
HISPANIC	3,729	1.99	1,680	0.57
INDIAN	33,742	18.00	19,043	6.47
ISRAELI	9,699	5.17	460	0.16
ITALIAN	9,598	5.12	283	0.10
JAPANESE	1,588	0.85	580	0.20
KOREAN	5,284	2.82	32,474	11.03
Total	187,427	100	294,324	100

Table 6: Distributions of Positive and Negative Name Instance Pairs per ENG in AMINER Training Data

ENG	Positive Pairs	Ratio (%)	Negative Pairs	Ratio (%)
CHINESE	38,958	61.88	76,781	91.26
ENGLISH	13,758	21.85	2,414	2.87
GERMAN	2,603	4.13	748	0.89
HISPANIC	1,933	3.07	380	0.45
INDIAN	5,701	9.06	3,815	4.53
Total	62,953	100	84,138	100

Table 7: Distributions of Positive and Negative Name Instance Pairs per ENG in UM-IRIS Training Data

ENG	Positive Pairs	Ratio (%)	Negative Pairs	Ratio (%)
CHINESE	1,517	10.60	7,261	37.54
ENGLISH	2,185	15.26	459	2.37
GERMAN	3,673	25.66	1,917	9.91
HISPANIC	2,416	16.88	970	5.01
INDIAN	2,080	14.53	3,803	19.66
KOREAN	2,445	17.08	4,933	25.50
Total	14,316	100	19,343	100

As noted above for Figure 5 and observed in other labeled data (see Figure S1 ~ S16 in Supplementary Material), the algorithms work better in finding more true negative pairs even for non-CHINESE name pairs when they are trained on data in which ethnic name partitioning is not performed (‘BEFORE’) and, thus, negative pairs are dominated by CHINESE ones as shown in Table 4~6. This implies that the nonmatch patterns in CHINESE name pairs are applicable to predicting nonmatch pairs for other ENGs. In contrast, during ENG-aware disambiguation, the algorithms come to rely on the small-size negative pairs that may skew or distort true nonmatch patterns for some ENGs. This seems to result in the decreased recall in predicting negative pairs (i.e., many true negatives classified as false positives, which reduces precision for positive pair prediction), while increasing slightly precision in predicting negative pairs.

Despite the aforementioned conflicting changes in precision and recall per ENG, the overall performance by the four algorithms on the whole test set are shown in Figure 1~4 to substantially increase across the four labeled data after ethnic name partitioning is included in machine learning. One reason would be that performance gains outweigh losses at each ENG level overall. Another reason would be that especially for KISTI, AMINER, and GESIS, the improved performances in disambiguating CHINESE that constitute the majority of name instances may affect the overall evaluation results. As shown by the case of UM-IRIS in which ENG sizes are controlled to be equal, however, the overall performance improvements can be observed for all the ENGs by ENG-aware disambiguation. As such, this study illustrates that the ethnic name partition can be truly effective in improving disambiguation performances.

Discussion

These results suggest that AND tasks may produce better results by using ethnic name partition in machine learning. Considering that adding more features can improve generally machine learning performances, the enhanced disambiguation performances by ENG partitioning might not be a surprise. With that said, the real contribution of this study would be that it demonstrates many machine learning based disambiguation models have a potential to be improved by introducing ethnic name grouping into ambiguous data without additional collection of feature information.

To fully realize this potential, however, a few issues need to be addressed. First, ENG tagging can be a non-trivial task that requires an intricate algorithmic technique itself. Thanks to the ENG classification system developed and publicly shared by Torvik and Agarwal (2016), this study could assign ENGs to the names in four labeled data. Although *Ethnea* was constructed based on more than 9 million author name instances in PubMed, the world largest biomedicine library, it is unknown how well it can help us tag ENGs to names in other fields. Ideally, *Ethnea* may be updated regularly to reflect new author names entering bibliographic data in various fields. Practically, further research may be focused on finding out a set of ENGs that are most influential in improving disambiguation results and thus simplifying ENG tagging for author name disambiguation (e.g., CHINESE vs Non-CHINESE).

Second, the findings of this study were based on three labeled data (KISTI, AMINER, and GESIS) in which CHINESE names are dominant and the overall performance improvements were heavily affected by those for CHINESE name instances. To overcome such an imbalance of instance distribution in labeled data, a new labeled data (UM-IRIS) were created in a way that six ENGs have the same amount of ambiguous name instances. Disambiguation results from the new labeled data were in line with those from other three labeled data. In addition, all ENGs including CHINESE were able to obtain gains in disambiguation performances. But all these findings were obtained from small-sized labeled data, whether they are biased or controlled for ENG sizes. So, it is still unknown whether such improvements are achievable in author name disambiguation for large-scale bibliographic data in which ENG composition may be quite different from those in the labeled data used in this paper.

Another issue would be that there can be other features than the four used in this study that can lead ethnic name partition to different AND performances. For example, English authors may appear in publication records that are more complete in affiliation information and use more diverse title terms. Meanwhile, Chinese authors may tend to work with coauthors who have similar names in same institutions. Various features need to be explored to study further the impact of ethnic name partition on AND.

Fourth, ENG-aware disambiguation may be beneficial for positive pair prediction but not so much for negative pair prediction. This was illustrated in Figure 5 above and Figure S1 ~ S16 in Supplementary Material by the dramatically decreased recall in negative pair predictions for many ENGs. It was contrasted with the substantial increase of precision in positive pair prediction for those ENGs. This study speculates that by ethnic name partitioning, classifiers become stricter for CHINESE pairs while relaxed for other ENG pairs. In other words, a pairs of CHINESE instances that would be classified as ‘match’ before partitioning are classified as ‘nonmatch’ after partitioning (Prec-Pos↑; Rec-Pos↓), while ‘nonmatch’ pairs of other ENG instances as matched ones (Prec-Pos↓; Rec-Pos↑). This might be because while some CHINESE pairs sharing coauthor names, venue names, or title words refer to different authors, other ethnic names sharing the features are more likely to represent the same authors (see Figure 6 B, D, F, and H in which Chinese name pair share is denoted in square). During training, such different similarity patterns are mixed up before partitioning but distinguished after it. Another conjecture is that due to the relaxed classification after partitioning, true negative pairs of other-than-CHINESE ENGs are falsely classified as false positive pairs (Prec-Pos↓; Rec-Neg↓). As the sizes of negative pairs in most non-CHINESE ENGs are smaller than positive pairs (see Table 4~7), misclassified negative pairs have larger impacts on Rec-Neg than on Prec-Pos across the ENGs. But these conjectures are based on the observations on labeled data in which Chinese name instances are prevalent. Using an additional labeled data with controlled ENG sizes, however, the conjecture has been confirmed. But only six ENGs in a small dataset were considered for analysis. More ENGs need to be investigated to check if this conjecture holds good under the different combinations of ENGs.

Conclusion and Discussion

This study evaluated the effects ethnic name partitioning has on author name disambiguation (AND) using machine learning methods. For this, author name instances in four labeled datasets were disambiguated under two scenarios. First, similarity scores of instance pairs over four basic features – author name, coauthor names, paper title, and publication venue – were used to train and test disambiguation algorithms. Second, in addition to the basic features, ethnic name group (ENGs) were tagged to name instances to allow algorithms to build models that are optimized to each ENG. Comparisons of disambiguation performances before and after ENG-aware disambiguation showed that using ethnic name partition can substantially improve algorithmic performances. Such performance improvements occurred across all ENGs, although performance gains and losses at each ENG level were observed in different ways depending on the types of measures – precision or recall – and target classifications – positive (match) or negative (nonmatch) pairs.

As detailed in the discussion above, ethnic name partition requires further research for us to understand better its impact on author name disambiguation and apply it to disambiguation tasks for digital libraries that are struggling with authority control over fast-growing ambiguous author names. This study is expected to motivate scholars and practitioners to study toward that direction by demonstrating the potential of ENG-aware disambiguation in improving disambiguation performances.

Acknowledgements

This paper was supported by grants from the National Science Foundation (#1917663, #1760609, and #1561687). The authors would like to thank anonymous reviewers for their comments to improve the paper.

References

Chin, A. W.-S., Juan, Y.-C., Zhuang, Y., Wu, F., Tung, H.-Y., Yu, T., . . . et al. (2013). *Effective string processing and matching for author disambiguation*. Paper presented at the Proceedings of the 2013 KDD Cup 2013 Workshop, Chicago, Illinois. https://doi.org/10.1145/2517288.2517295

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. F. (2010). An Unsupervised Heuristic-Based Hierarchical Method for Name Disambiguation in Bibliographic Citations. *Journal of the American Society for Information Science and Technology, 61*(9), 1853-1870. doi:10.1002/asi.21363

Deville, P., Wang, D. S., Sinatra, R., Song, C. M., Blondel, V. D., & Barabási, A.-L. (2014). Career on the Move: Geography, Stratification, and Scientific Impact. *Scientific Reports, 4*. doi:10.1038/srep04770

Fegley, B. D., & Torvik, V. I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLOS ONE, 8*(7). doi:10.1371/journal.pone.0070299

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. *Sigmod Record, 41*(2), 15-26.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. F. (2014). Self-Training Author Name Disambiguation for Information Scarce Scenarios. *Journal of the Association for Information Science and Technology, 65*(6), 1257-1278. doi:10.1002/asi.22992

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). *Two Supervised Learning Approaches for Name Disambiguation in Author Citations.* Paper presented at the Jcdl 2004: Proceedings of the Fourth Acm/Ieee Joint Conference on Digital Libraries, Tucson, Arizona.

Han, H., Zha, H. Y., & Giles, C. L. (2005). Name disambiguation in author citations using a K-way spectral clustering method. *Proceedings of the 5th Acm/Ieee Joint Conference on Digital Libraries, Proceedings*, 334-343. doi:Doi 10.1145/1065385.1065462

Harzing, A. W. (2015). Health warning: might contain multiple personalities-the problem of homonyms in Thomson Reuters Essential Science Indicators. *Scientometrics, 105*(3), 2259-2270. doi:10.1007/s11192-015-1699-y

Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review, 32*, e22. doi:10.1017/S0269888917000182

Hussain, I., & Asghar, S. (2018). DISC: Disambiguating homonyms using graph structural clustering. *Journal of Information Science, 44*(6), 830-847. doi:10.1177/0165551518761011

Islamaj Dogan, R., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database : the journal of biological databases and curation, 2009*, bap018-bap018. doi:10.1093/database/bap018

Kang, I. S., Kim, P., Lee, S., Jung, H., & You, B. J. (2011). Construction of a Large-Scale Test Set for Author Disambiguation. *Information Processing & Management, 47*(3), 452-465. doi:10.1016/j.ipm.2010.10.001

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics, 103*(3), 1061-1071. doi:10.1007/s11192-015-1580-z

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: A case of DBLP. *Scientometrics, 116*(3), 1867-1886. doi:10.1007/s11192-018-2824-5

Kim, J., & Diesner, J. (2015). The effect of data pre-processing on understanding the evolution of collaboration networks. *Journal of Informetrics, 9*(1), 226-236. doi:10.1016/j.joi.2015.01.002

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology, 67*(6), 1446-1461. doi:10.1002/asi.23489

Kim, J., & Owen-Smith, J. (In print). ORCID-linked labeled data for evaluating author name disambiguation at scale. Scientometrics. doi:10.1007/s11192-020-03826-6

Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics, 117*(1), 511-526. doi:10.1007/s11192-018-2865-9

Kim, J., & Kim, J. (2020). Effect of Forename String on Author Name Disambiguation. *Journal of the Association for Information Science and Technology, 71*(7), 839-855. doi:10.1002/asi.24298

Kim, K., Sefid, A., & Giles, C. L. (2017). Scaling Author Name Disambiguation with CNF Blocking. *arXiv preprint arXiv:1709.09657*.

Kim, K., Sefid, A., Weinberg, B. A., & Giles, C. L. (2018). *A Web Service for Author Name Disambiguation in Scholarly Databases.* Paper presented at the 2018 IEEE International Conference on Web Services (ICWS).

Lerchenmueller, M. J., & Sorenson, O. (2016). Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists. *PLOS ONE, 11*(7), e0158731. doi:10.1371/journal.pone.0158731

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-Based Bootstrapping for Large-Scale Author Disambiguation. *Journal of the American Society for Information Science and Technology, 63*(5), 1030-1047. doi:10.1002/asi.22621

Ley, M. (2009). DBLP: some lessons learned. *Proceedings of the VLDB Endowment, 2*(2), 1493-1500.

Louppe, G., Al-Natsheh, H. T., Susik, M., & Maguire, E. J. (2016). ENG Sensitive Author Disambiguation Using Semi-supervised Learning. *Knowledge Engineering and Semantic Web, Kesw 2016, 649*, 272-287. doi:10.1007/978-3-319-45880-9_21

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics, 7*(4), 767-773. doi:10.1016/j.joi.2013.06.006

Momeni, F., & Mayr, P. (2016). *Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names.* Paper presented at the 20th international Conference on Theory and Practice of Digital Libraries (TPDL 2016), Hannover, Germany.

Müller, M. C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics, 111*(3), 1467-1500. doi:10.1007/s11192-017-2363-5

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2015). On the Combination of Domain-Specific Heuristics for Author Name Disambiguation: The Nearest Cluster Method. *International Journal on Digital Libraries, 16*(3-4), 229-246. doi:10.1007/s00799-015-0158-y

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., & Ferreira, A. A. (2017). Incremental author name disambiguation by exploiting domain-specific heuristics. *Journal of the Association for Information Science and Technology, 68*(4), 931-945. doi:10.1002/asi.23726

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2019). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*. doi:10.1177/0165551519888605

Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics, 107*(3), 1283-1298. doi:10.1007/s11192-016-1892-7

Shin, D., Kim, T., Choi, J., & Kim, J. (2014). Author Name Disambiguation Using a Graph Model with Node Splitting and Merging Based on Bibliographic Information. *Scientometrics, 100*(1), 15-50. doi:10.1007/s11192-014-1289-4

Smalheiser, N. R., & Torvik, V. I. (2009). Author Name Disambiguation. *Annual Review of Information Science and Technology, 43*, 287-313.

Song, M., Kim, E. H. J., & Kim, H. J. (2015). Exploring Author Name Disambiguation on PubMed-Scale. *Journal of Informetrics, 9*(4), 924-941. doi:10.1016/j.joi.2015.08.004

Strotmann, A., & Zhao, D. Z. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology, 63*(9), 1820-1833. doi:10.1002/asi.22695

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). *ArnetMiner: extraction and mining of academic social networks*. Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA.

Torvik, V. I., & Agarwal, S. (2016). *Ethnea: An instance-based ENG classifier based on geo-coded author names in a large-scale bibliographic database*. Paper presented at the Library of Congress International Symposium on Science of Science, Washington D.C. http://hdl.handle.net/2142/88927

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *Acm Transactions on Knowledge Discovery from Data, 3*(3). doi:10.1145/1552303.1552304

Treeratpituk, P., & Giles, C. L. (2009). Disambiguating Authors in Academic Publications using Random Forests. *JCDL 2009: Proceedings of the 2009 Acm/Ieee Joint Conference on Digital Libraries*, 39-48.

Vishnyakova, D., Rodriguez-Esteban, R., Ozol, K., & Rinaldi, F. (2016, dec). *Author Name Disambiguation in MEDLINE Based on Journal Descriptors and Semantic Types.* Paper presented at the Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), Osaka, Japan.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics, 93*(2), 391-411. doi:10.1007/s11192-012-0681-1

Wang, X., Tang, J., Cheng, H., & Yu, P. S. (2011). *ADANA: Active Name Disambiguation*. Paper presented at the 2011 IEEE 11th International Conference on Data Mining.

Wu, H., Li, B., Pei, Y. J., & He, J. (2014). Unsupervised author disambiguation using Dempster-Shafer theory. *Scientometrics, 101*(3), 1955-1972. doi:10.1007/s11192-014-1283-x

Wu, J., & Ding, X. H. (2013). Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics, 96*(3), 683-697. doi:10.1007/s11192-013-0978-8

Zhao, Z., Rollins, J., Bai, L., & Rosen, G. (2017, Oct. 2017). *Incremental Author Name Disambiguation for Scientific Citation Data.* Paper presented at the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Zhu, J., Wu, X., Lin, X., Huang, C., Fung, G. P. C., & Tang, Y. (2018). A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering. *Scientometrics, 114*(3), 781-794. doi:10.1007/s11192-017-2611-8

Table 1: Summary of Ethnic Name Group (ENG) Frequencies in Labeled Data

Labeled Data (No. of Instances)							
KISTI (41,605)		**AMINER (7,528)**		**GESIS (29,965)**		**UM-IRIS (6,000)**	
ENG	**Ratio (%)**	**ENG**	**Ratio (%)**	**ENG**	**Ratio (%)**	**ENG**	**Ratio (%)**
CHINESE	50.1	CHINESE	64.2	CHINESE	56.2	CHINESE	16.67
ENGLISH	15.6	ENGLISH	17.0	GERMAN	14.8	ENGLISH	16.67
INDIAN	9.5	INDIAN	6.3	ENGLISH	7.0	GERMAN	16.67
KOREAN	8.0	GERMAN	5.6	INDIAN	4.3	HISPANIC	16.67
GERMAN	3.3	HISPANIC	3.1	HISPANIC	3.6	INDIAN	16.67
ISRAELI	2.2	**Sum**	**96.2**	KOREAN	3.5	KOREAN	16.67
ITALIAN	2.0	Excluded ENGs (3.8%): JAPANESE, NORDIC, KOREAN, ARAB		JAPANESE	2.8	**Sum**	**100.00**
HISPANIC	1.7			ITALIAN	1.7	Excluded ENGs (0%): None	
JAPANESE	1.1			ARAB	1.6		
DUTCH	1.0			FRENCH	1.3		
ARAB	0.9			**Sum**	**96.8**		
FRENCH	0.9			Excluded ENGs (3.2%): NORDIC, HUNGARIAN, DUTCH, VIETNAMESE, SLAV, ROMANIAN, GREEK, ISRAELI, TURKISH, INDONESIAN			
Sum	**96.3**						
Excluded ENGs (3.7%) : NORDIC, SLAV, GREEK, ROMANIAN, VIETNAMESE, NULL, AFRICAN, TURKISH, HUNGARIAN,							

Table 2: A Mock-Up Example of Cosine Similarity Scores for Instance Pairs over Four Features and ENG Encoding

Pairs	Feature								Label
	Author Name	Coauthor Name	Paper Title	Pub. Venue	ENG 1	ENG 2	ENG 3	…	
Pair 1	0.97	0.89	0.67	0.12	Yes	No	No	**…**	Match
Pair 2	1.00	0.24	0.46	0.00	No	Yes	No	**…**	Nonmatch
Pair 3	0.65	0.07	0.00	0.80	No	No	Yes	**…**	Nonmatch
Pair 4	0.58	0.08	0.00	0.00	Yes	No	No	**…**	Nonmatch

Table 3: Numbers of Correctly or Incorrectly Predicted Pairs for Positive and Negative Pairs by Four Algorithms on KISTI Test Data

Algorithm	ENGs Considered	No. of Pairs (P: Positive) (N: Negative)	TP	FN	FP	TN
LR	Before	483,029 (P: 189,375) (N: 293,654)	55,998	133,377	20,203	273,451
	After		131,364	58,011	39,068	254,586
NB	Before		50,782	138,593	14,267	279,387
	After		71,020	118,355	8,775	284,879
RF	Before		82,727	106,648	57,532	236,122
	After		128,105	61,270	40,479	253,175
GB	Before		75,195	114,180	24,725	268,929
	After		136,859	52,516	39,888	253,766

Table 4: Distributions of Positive and Negative Name Instance Pairs per ENG in GESIS Training Data

ENG	Positive Pairs	Ratios	Negative Pairs	Ratios
ARAB	1,559	0.66	1,203	0.21
CHINESE	145,969	62.22	551,150	93.96
ENGLISH	16,020	6.83	2,147	0.37
FRENCH	1,948	0.83	3,118	0.53
GERMAN	37,373	15.93	13,092	2.23
HISPANIC	5,887	2.51	1,907	0.33
INDIAN	6,708	2.86	2,654	0.45
ITALIAN	3,489	1.49	686	0.12
JAPANESE	8,853	3.77	2,608	0.44
KOREAN	6,792	2.90	7,994	1.36
Total	234,598	100	586,559	100

Table 5: Distributions of Positive and Negative Name Instance Pairs per ENG in KISTI Training Data

ENG	Positive Pairs	Ratios	Negative Pairs	Ratios
ARAB	6,015	3.21	37	0.01
CHINESE	53,519	28.55	209,807	71.28
DUTCH	5,448	2.91	84	0.03
ENGLISH	45,054	24.04	25,137	8.54
FRENCH	2,686	1.43	625	0.21
GERMAN	11,065	5.90	4,114	1.40
HISPANIC	3,729	1.99	1,680	0.57
INDIAN	33,742	18.00	19,043	6.47
ISRAELI	9,699	5.17	460	0.16
ITALIAN	9,598	5.12	283	0.10
JAPANESE	1,588	0.85	580	0.20
KOREAN	5,284	2.82	32,474	11.03
Total	187,427	100	294,324	100

Table 6: Distributions of Positive and Negative Name Instance Pairs per ENG in AMINER Training Data

ENG	Positive Pairs	Ratio (%)	Negative Pairs	Ratio (%)
CHINESE	38,958	61.88	76,781	91.26
ENGLISH	13,758	21.85	2,414	2.87
GERMAN	2,603	4.13	748	0.89

HISPANIC	1,933	3.07	380	0.45
INDIAN	5,701	9.06	3,815	4.53
Total	62,953	100	84,138	100

Table 7: Distributions of Positive and Negative Name Instance Pairs per ENG in UM-IRIS Training Data

ENG	**Positive Pairs**	**Ratio (%)**	**Negative Pairs**	**Ratio (%)**
CHINESE	1,517	10.60	7,261	37.54
ENGLISH	2,185	15.26	459	2.37
GERMAN	3,673	25.66	1,917	9.91
HISPANIC	2,416	16.88	970	5.01
INDIAN	2,080	14.53	3,803	19.66
KOREAN	2,445	17.08	4,933	25.50
Total	14,316	100	19,343	100

Table 1: Summary of Ethnic Name Group (ENG) Frequencies in Labeled Data

Labeled Data (No. of Instances)							
KISTI (41,605)		**AMINER (7,528)**		**GESIS (29,965)**		**UM-IRIS (6,000)**	
ENG	**Ratio (%)**	**ENG**	**Ratio (%)**	**ENG**	**Ratio (%)**	**ENG**	**Ratio (%)**
CHINESE	50.1	CHINESE	64.2	CHINESE	56.2	CHINESE	16.67
ENGLISH	15.6	ENGLISH	17.0	GERMAN	14.8	ENGLISH	16.67
INDIAN	9.5	INDIAN	6.3	ENGLISH	7.0	GERMAN	16.67
KOREAN	8.0	GERMAN	5.6	INDIAN	4.3	HISPANIC	16.67
GERMAN	3.3	HISPANIC	3.1	HISPANIC	3.6	INDIAN	16.67
ISRAELI	2.2	**Sum**	**96.2**	KOREAN	3.5	KOREAN	16.67
ITALIAN	2.0	Excluded ENGs (3.8%): JAPANESE, NORDIC, KOREAN, ARAB		JAPANESE	2.8	**Sum**	**100.00**
HISPANIC	1.7			ITALIAN	1.7	Excluded ENGs (0%): None	
JAPANESE	1.1			ARAB	1.6		
DUTCH	1.0			FRENCH	1.3		
ARAB	0.9			**Sum**	**96.8**		
FRENCH	0.9			Excluded ENGs (3.2%): NORDIC, HUNGARIAN, DUTCH, VIETNAMESE, SLAV, ROMANIAN, GREEK, ISRAELI, TURKISH, INDONESIAN			
Sum	**96.3**						
Excluded ENGs (3.7%) : NORDIC, SLAV, GREEK, ROMANIAN, VIETNAMESE, NULL, AFRICAN, TURKISH, HUNGARIAN,							

Table 2: A Mock-Up Example of Cosine Similarity Scores for Instance Pairs over Four Features and ENG Encoding

Pairs	Feature							Label
	Author Name	Coauthor Name	Paper Title	Pub. Venue	ENG 1	ENG 2	ENG 3	…	
Pair 1	0.97	0.89	0.67	0.12	Yes	No	No	…	Match
Pair 2	1.00	0.24	0.46	0.00	No	Yes	No	…	Nonmatch
Pair 3	0.65	0.07	0.00	0.80	No	No	Yes	…	Nonmatch
Pair 4	0.58	0.08	0.00	0.00	Yes	No	No	…	Nonmatch

Table 3: Numbers of Correctly or Incorrectly Predicted Pairs for Positive and Negative Pairs by Four Algorithms on KISTI Test Data

Algorithm	ENGs Considered	No. of Pairs (P: Positive) (N: Negative)	TP	FN	FP	TN
LR	Before	483,029 (P: 189,375) (N: 293,654)	55,998	133,377	20,203	273,451
	After		131,364	58,011	39,068	254,586
NB	Before		50,782	138,593	14,267	279,387
	After		71,020	118,355	8,775	284,879
RF	Before		82,727	106,648	57,532	236,122
	After		128,105	61,270	40,479	253,175
GB	Before		75,195	114,180	24,725	268,929
	After		136,859	52,516	39,888	253,766

Table 4: Distributions of Positive and Negative Name Instance Pairs per ENG in GESIS Training Data

ENG	Positive Pairs	Ratios	Negative Pairs	Ratios
ARAB	1,559	0.66	1,203	0.21
CHINESE	145,969	62.22	551,150	93.96
ENGLISH	16,020	6.83	2,147	0.37
FRENCH	1,948	0.83	3,118	0.53
GERMAN	37,373	15.93	13,092	2.23
HISPANIC	5,887	2.51	1,907	0.33
INDIAN	6,708	2.86	2,654	0.45
ITALIAN	3,489	1.49	686	0.12
JAPANESE	8,853	3.77	2,608	0.44
KOREAN	6,792	2.90	7,994	1.36
Total	234,598	100	586,559	100

Table 5: Distributions of Positive and Negative Name Instance Pairs per ENG in KISTI Training Data

ENG	Positive Pairs	Ratios	Negative Pairs	Ratios
ARAB	6,015	3.21	37	0.01
CHINESE	53,519	28.55	209,807	71.28
DUTCH	5,448	2.91	84	0.03
ENGLISH	45,054	24.04	25,137	8.54
FRENCH	2,686	1.43	625	0.21
GERMAN	11,065	5.90	4,114	1.40
HISPANIC	3,729	1.99	1,680	0.57
INDIAN	33,742	18.00	19,043	6.47
ISRAELI	9,699	5.17	460	0.16
ITALIAN	9,598	5.12	283	0.10
JAPANESE	1,588	0.85	580	0.20
KOREAN	5,284	2.82	32,474	11.03
Total	187,427	100	294,324	100

Table 6: Distributions of Positive and Negative Name Instance Pairs per ENG in AMINER Training Data

ENG	Positive Pairs	Ratio (%)	Negative Pairs	Ratio (%)
CHINESE	38,958	61.88	76,781	91.26
ENGLISH	13,758	21.85	2,414	2.87
GERMAN	2,603	4.13	748	0.89

HISPANIC	1,933	3.07	380	0.45
INDIAN	5,701	9.06	3,815	4.53
Total	62,953	100	84,138	100

Table 7: Distributions of Positive and Negative Name Instance Pairs per ENG in UM-IRIS Training Data

ENG	**Positive Pairs**	**Ratio (%)**	**Negative Pairs**	**Ratio (%)**
CHINESE	1,517	10.60	7,261	37.54
ENGLISH	2,185	15.26	459	2.37
GERMAN	3,673	25.66	1,917	9.91
HISPANIC	2,416	16.88	970	5.01
INDIAN	2,080	14.53	3,803	19.66
KOREAN	2,445	17.08	4,933	25.50
Total	14,316	100	19,343	100