Improving Machine-based Entity Resolution with Limited Human Effort: A Risk Perspective

Zhaoqiang Chen, Qun Chen, Boyi Hou, Murtadha Ahmed, Zhanhuai Li

May 31, 2018

1 ABSTRACT

Pure machine-based solutions usually struggle in the challenging classification tasks such as entity resolution (ER). To alleviate this problem, a recent trend is to involve the human in the resolution process, most notably the crowdsourcing approach. However, it remains very challenging to effectively improve machine-based entity resolution with limited human effort. In this paper, we investigate the problem of human and machine cooperation for ER from a risk perspective. We propose to select the machine-labeled instances at high risk of being mislabeled for manual verification. For this task, we present a risk model that takes into consideration the human-labeled instances as well as the output of machine resolution. Finally, we evaluate the performance of the proposed risk model on real data. Our experiments demonstrate that it can pick up the mislabeled instances with considerably higher accuracy than the existing alternatives. Provided with the same amount of human cost budget, it can also achieve better resolution quality than the state-of-the-art approach based on active learning.

2 INTRODUCTION

Entity resolution aims at finding the records that refer to the same real-world entity. Usually considered as a classification task [1], ER is challenging in that the records may

contain incomplete and dirty values. ER can be performed based on rules, probabilistic theory or machine learning [1, 2]. However, the traditional machine-based solutions may not be able to produce satisfactory results in many practical scenarios. Therefore, there is an increasing need to involve the human in the resolution process for improved quality [3]. For instance, the active learning approach [4] proposed to select the instances for manual verification based on the benefit they can bring to a machine classifier. The approach of crowdsourcing [3, 5, 6] instead investigated how to make the human work efficiently and effectively on a given workload. Depending on pre-specified assumptions (e.g. partial order relationship [7]), it usually makes the human label some instances in a workload for the purpose that the remaining instances can be automatically labeled by the machine with high accuracy.

It can be observed that the existing hybrid approaches select the instances for manual verification to maximize the benefit they can bring to a given workload as a whole. However, the marginal benefit of additional manual work usually decreases (sometimes dramatically) with the cost. For instance, in active learning, it has been well recognized [8] that increasing the number of training data points may quickly become ineffectual in improving classification performance after initial iterations. In the application scenarios where fast response is required, it is also desirable that a limited amount of human effort can be exclusively spent on the instances at high risk of being mislabeled by the machine.

In this paper, we investigate the problem of human and machine cooperation for improved quality from a risk perspective. Given a limited human cost budget, we propose to select the machine-labeled instances at high risk of being mislabeled for manual verification. The proposed risk-based solution is supposed to be used in the scenario where increasing training points for a learning model has become ineffectual or not cost-effective in improving classification performance. It can therefore serve as a valuable complement to the existing learning-based solutions. On the other hand, even though some of the proposed techniques for active learning (e.g. training instance selection based on uncertainty [9]) can be naturally applied for this task, our work is the first to introduce the concept of risk and propose a formal risk model for the task. The major contributions of this paper can be summarized as follows:

- We investigate the problem of human and machine cooperation for ER from a risk perspective and define the corresponding optimization problem (Section. 3);

- We present a risk model for prioritizing the machine-labeled instances for manual verification (Section. 4);

- We evaluate the performance of the proposed approach on real data by a comparative study. The experimental results validate its efficacy (Section. 5).

3 Problem Definition

Given an ER workload consisting of record pairs, a machine classifier labels each pair as *match* or *unmatch*. Due to the inherent challenge of entity resolution, a classifier may
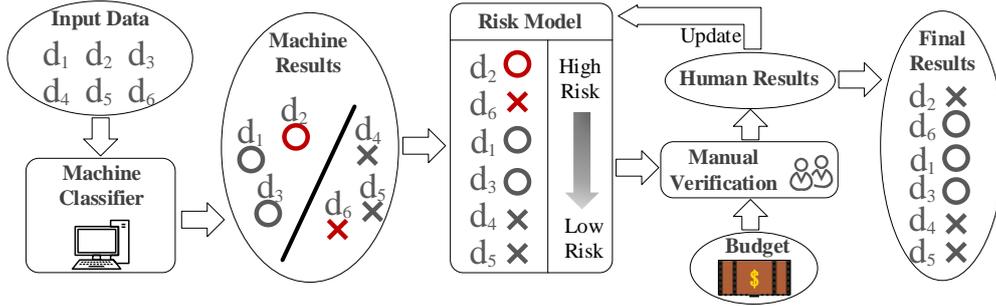
Figure 3.1: The Risk-based Solution.

be prone to mislabeling some of the pairs. In this paper, we investigate the problem of how to improve the results of machine resolution by manually correcting machine errors. Since human work is expensive, we impose a budget on the amount of spent human effort. For the sake of presentation simplicity, we quantify the budget by the number of manually-inspected pairs. Given a budget k, an ideal solution would identify k mislabeled pairs. In this case, each manual inspection effectively corrects a machine error. However, in practice, it is more likely that a solution chooses both mislabeled and correctly labeled pairs. We formally define the optimization problem as follows:

Definition 1 *[Optimization Problem of Improving Machine Resolution by Manual Inspection]. Given an ER workload, D, which consists of n record pairs, $\{d_1, d_2, \ldots, d_n\}$, a machine classifier labels each pair in D as* match *or* unmatch. *Given the budget k on human work, the optimization problem is to identify a set of k machine-labeled pairs in D, denoted by D_H, for manual inspection such that the number of pairs misclassified by the machine in D_H is maximized.*

Risk-based Solution. The optimization problem defined in Definition. 1 is challenging due to the fact that the match probabilities of the machine-labeled pairs are difficult to estimate. In this paper, we propose to solve the optimization problem from a risk perspective. In other words, the machine-labeled pairs at higher risk of being mislabeled should be chosen first for manual inspection. It can be observed that if risk measurement is accurate given all the available information, the strategy of selecting by risk-wise order can be considered optimal. The workflow of the risk-based solution is presented in Figure. 3.1. It iteratively selects the most risky machine-labeled pairs for manual inspection until the budget limit is reached. After each iteration, the set of manually-labeled pairs is updated, and is used to re-evaluate the risk of the remaining machine-labeled pairs.

It is worthy to point out that the risk-based solution can work properly with both supervised and unsupervised classifiers. Given a supervised classifier, risk analysis can be initially performed based on the human-labeled pairs as well as machine resolution. Given an unsupervised classifier, risk analysis can only start with machine resolution; after initial iterations, it can then be similarly performed based on the human-labeled pairs as well as machine resolution.

3

4 RISK ANALYSIS

In this section, we propose the technique of risk analysis for prioritizing pair selection. Given an instance pair d_i in D, we represent its match probability by a random variable, P_i. As usual, we model P_i by a normal distribution, $\mathcal{N}(\mu_i, \sigma_i^2)$, where μ_i and σ_i^2 denote its expectation and variance respectively. In the rest of this section, we first describe how to estimate the match probability distribution in Subsection 4.1, and then present the metric for risk measurement in Subsection 4.2.

4.1 DISTRIBUTION ESTIMATION

It can be observed that there exist two information sources for the estimation of match probability distribution. Firstly, even though a machine classifier may fail to produce satisfactory resolution results, it can provide valuable hints about the status of the pairs. Therefore, the results of machine resolution can generally serve as a starting point for the estimation. The second source consists of the human-labeled results. Compared with machine labels, the labels provided by the human are usually more accurate, i.e. they can provide more information beyond the capability of machine resolution.

We employ the classical Bayesian inference [10] to estimate the distribution. The inference process takes the match probability estimated by the machine as the prior expectation, and uses the human-labeled pairs as samples to estimate the posterior expectation and variance. The proposed approach has the desirable property that it can seamlessly integrate the hints provided by both the human and the machine into a unified inference process.

4.1.1 PRIOR EXPECTATION ESTIMATION BY MACHINE

A machine classifier labels instance pairs as match or unmatch based on a classification metric. Generally, the match probability of a pair can be considered to be monotonous with its metric value. In this paper, we use the SVM (Support Vector Machine) classifier based on active learning as the illustrative example. It classifies pairs through a hyperplane. Instead of randomly selecting training data points, it iteratively chooses the instance pair which is closest to the hyperplane of the current SVM as the next training data point, and updates the SVM until a preset training budget is exhausted. Note that an SVM classifier usually provides a pair's distance from the hyperplane, rather than a match probability, as the evidence for its given label. We therefore use Platt's probabilistic outputs for SVM [11] to translate the distance into a match probability.

4.1.2 SAMPLE OBSERVATION GENERATION BY HUMAN

We generate the sample observations on the status of a target pair based on features. Features serve as the medium to convey valuable information from the human-labeled pairs to a target pair. Desirably, the features used for information conveyance should have the following three properties:

1. They can be easily extracted from the human-labeled pairs;

2. They should be evidential, or indicative of the status of a pair;

3. They should be to a large extent independent of the metric used by the machine classifier.

The final property ensures that the sample observations can provide additional valuable information not implied by machine labels. To this aim, we extract two types of features from pairs, *Same(t_i)* and *Diff(t_i)*, where t_i represents a token, *Same(t_i)* indicates that both records in a pair contain t_i, and *Diff(t_i)* indicates that one and only one record in a pair contains t_i. It can be observed that these two features are evidential and easily extractable. Moreover, they were not used in the existing classification metrics proposed for ER.

Suppose that a target pair, d_i, contains m features, which are denoted by $\{f_1, f_2, \ldots, f_m\}$. A human-labeled pair containing all the m features can be naturally considered to be a valid observation on the status of d_i. Unfortunately, due to their limited number in practical scenarios, the human-labeled pairs with this property may not provide with sufficient observations. Therefore, we also consider the human-labeled pairs that contain only a portion of the m features in d_i. Suppose that a human-labeled pair, d_j^h, contains the k features in d_i, $\{f_1, f_2, \ldots, f_k\}$, but does not contain the remaining $(m - k)$ features. Inspired by portfolio investment theory [12], we treat features as stocks, and a feature's match probability as its investment reward. Then, the match probability of d_i corresponds to the combined reward of an investment portfolio consisting of m stocks, $\{f_1, f_2, \ldots, f_m\}$.

Based on the label of d_j^h, we generate the corresponding sample observation on the status of d_i by

$$O_j(d_i) = \frac{L(d_j^h) + \sum_{k < r \leq m} w_r E(f_r)}{1 + \sum_{k < r \leq m} w_r}, \tag{4.1}$$

in which w_r denotes the feature weight, $L(d_j^h)$ denotes the manual label of d_j^h, and $E(f_r)$ denotes the expectation of f_r's match probability. In Eq. 4.1, $L(d_j^h){=}1$ if the label is *match* and $L(d_j^h){=}0$ otherwise. We estimate $E(f_r)$ by

$$E(f_r) = \frac{\sum_{1 \leq s \leq n} L(d_s^r)}{n}, \tag{4.2}$$

in which d_s^r denotes a human-labeled pair containing the feature f_r and n denotes its total number. An example of sample observation generation is shown in Example 1. To determine the feature weight w_r, we take two factors into account: the *informativeness* and *discriminability* of a feature. The *informativeness* describes the extent that a feature can help to determine the matching status. It is defined as $IF(f_r) = max(\frac{E(f_r)}{1-E(f_r)}, \frac{1-E(f_r)}{E(f_r)})$. And the *discriminability* describes the extent that a feature can identify the target pair. It is defined as *inverse document frequency* (IDF). Combining them together, we define $w_r = IF \cdot IDF$. It is worthy to point out that in the generation of sample observations for d_i, we only consider the features contained in the human-labeled pairs. If a feature

of d_i never appears in the human-labeled pairs, we lack reliable information to reason about its match probability. It is therefore ignored in the observation generation process.

Example 1 *Suppose that a target pair, d_1, contains 3 features, $\{f_1, f_2, f_3\}$, and a pair manually labeled as* unmatch *by the human, d_2^h, contains f_1 and f_2, but not f_3. For the sake of presentation simplicity, we also suppose that feature weights are equally set to be 1. With the expectation of the match probability of f_3 being estimated at 0.3, the sample observation provided by d_2^h for the status of d_1 is approximated by $O_2(d_1) = \frac{0+0.3}{2} = 0.15$.*

4.1.3 Bayesian Inference

Given a random variable V following a known prior distribution, $\pi(V)$, the technique of Bayesian inference [10] estimates the posterior distribution of V by combining the prior information provided by $\pi(V)$ and the sample observations. In our example, the prior distribution of the match probability of a target pair, d_i, is represented by the normal distribution of $\mathcal{N}(\mu_i, \sigma_i^2)$. Suppose that the prior expectation of μ_i provided by the machine classifier is μ_i^0 and the human-labeled pairs provide with n sample observations.

As usual, we suppose that μ_i and σ_i^2 follow a combined conjugate prior distribution, or a *normal-inverse-gamma distribution* [13]. The prior distributions of μ_i and σ_i^2 can thus be represented by

$$p(\mu_i | \sigma_i^2; \mu_i^0, n^0) \sim \mathcal{N}(\mu_i^0, \frac{\sigma_i^2}{n^0}), \tag{4.3}$$

and

$$p(\sigma_i^2; \alpha, \beta) \sim InvGamma(\alpha^0, \beta^0), \tag{4.4}$$

where n^0, α^0 and β^0 are the hyperparameters, and $InvGamma()$ denotes an *inverse-gamma distribution* [14]. Denoting the posteriors by $\mathcal{N}(\mu_i^1, \frac{\sigma_i^2}{n^1})$ and $InvGamma(\alpha^1, \beta^1)$, we have

$$
\begin{aligned}
\mu_i^1 &= \frac{n^0 \cdot \mu_i^0 + n \cdot \bar{p}_i}{n^0 + n}, \\
n^1 &= n^0 + n, \\
\alpha^1 &= \alpha^0 + \frac{n}{2}, \\
\beta^1 &= \beta^0 + \frac{1}{2} \sum_{j=1}^n (p_i^j - \bar{p}_i)^2 + \frac{1}{2} \cdot \frac{n^0 n}{n^0 + n} \cdot (\mu_i^0 - \bar{p}_i)^2,
\end{aligned}
\tag{4.5}
$$

where $\bar{p}_i$ denotes the average value of observed samples.

In Eq. 4.3 and 4.4, the hyperparameters n^0, α^0 and β^0 are used to convey the belief about the prior information. Specifically, given a confidence level θ on the prior expectation μ_i^0, we set $n^0 = \theta n/(1-\theta)$. It means that the inference process will preserve $\theta \mu_i^0$ for the estimation of μ_i. Similarly, we set $\alpha^0 = \frac{n}{2} \cdot \frac{\theta}{1-\theta} + 1$, and $\beta^0 = S_n^2 \cdot (\alpha^0 - 1)$, in which S_n^2 represents the variance of all the samples. It means that the inference process will preserve θS_n^2 for the estimation of σ_i^2.

Based on the obtained posterior distributions of μ_i and σ_i^2, a point estimate $\hat{\mu}_i$ for the random variable μ_i (resp. $\hat{\sigma}_i^2$ for σ_i^2) can be inferred using a metric of Bayes risk. A

common risk function is the mean square error, which is defined as $E[(\hat{\chi}-\chi)^2]$. Using the mean of the posterior distribution as the Bayes estimate of the unknown parameter can minimize the mean square error [10]. Having known that the mean of an inverse-gamma distribution with parameters α and β is $\frac{\beta}{\alpha-1}$, the Bayes estimates of the expectation and the variance can be represented as follows,

$$\hat{\mu}_i = \mu_i^1, \tag{4.6}$$

$$\hat{\sigma}_i^2 = \frac{\beta^1}{\alpha^1 - 1}. \tag{4.7}$$

4.2 Risk Model

Inspired by portfolio investment theory [12], we employ the metric of Conditional Value at Risk (CVaR) to measure the risk of pairs being mislabeled by the machine. Given a confidence level of θ, CVaR is the expected loss incurred in the $1 - \theta$ worst cases. Formally, given the loss function $z(X) \in L^p(\mathcal{F})$ of a portfolio X and θ, the metric of CVaR is defined as follows:

$$CVaR_\theta(X) = \frac{1}{1-\theta} \int_0^{1-\theta} VaR_{1-\gamma}(X)d\gamma, \tag{4.8}$$

where $VaR_{1-\gamma}(X)$ represents the minimum loss incurred at or below γ and can be formally represented by

$$VaR_{1-\gamma}(X) = inf\{z_* : P(z(X) \geq z_*) \leq \gamma\}. \tag{4.9}$$

Given a pair, d_i, we denote its match probability by x, and its probability density function and cumulative distribution function by $pdf_{d_i}(x)$ and $cdf_{d_i}(x)$ respectively. If d_i is labeled by the machine as *unmatch*, its probability of being mislabeled by the machine is equal to x. Accordingly, its worst-case loss corresponds to the case that x is maximal. Therefore, given the confidence level of θ, the CVaR of d_i is the expectation of $z = x$ in the $1 - \theta$ cases where x is from $cdf_{d_i}^{-1}(\theta)$ to $+\infty$. Formally, the CVaR risk of a pair d_i with the machine label of *unmatch* can be estimated by

$$CVaR_\theta(d_i) = \frac{1}{1-\theta} \int\limits_{cdf_{d_i}^{-1}(\theta)}^{+\infty} pdf_{d_i}(x) \cdot x dx. \tag{4.10}$$

Otherwise, if d_i is labeled by the machine as *match*, its potential loss of being mislabeled by the machine is equal to $1 - x$. Therefore, the CVaR risk of a pair d_i with the machine label of *match* can be similarly estimated by

$$CVaR_\theta(d_i) = \frac{1}{1-\theta} \int\limits_{-\infty}^{cdf_{d_i}^{-1}(1-\theta)} pdf_{d_i}(x) \cdot (1-x) dx. \tag{4.11}$$

5 Empirical Evaluation

We have evaluated the performance of the proposed risk model, denoted by CVAR, on real data by a comparative study. We compare it with both a baseline alternative and a state-of-the-art technique proposed for active learning [9]. The baseline method, denoted by BASE, selects the machine-labeled pairs solely based on the match expectation estimated by the machine. Specifically, given a pair d_i and its match probability μ_i^0 provided by a classifier, the risk of d_i with the machine label of *unmatch* (resp. *match*) is simply estimated to be μ_i^0 (resp. $(1 - \mu_i^0)$). Since the two algorithms proposed in [9], *Uncertainty* and *MinExpError*, perform very similarly in our experiments, we only report the results of *Uncertainty*. We denote the algorithm of *Uncertainty* by UNCT. Intuitively, UNCT iteratively selects the pairs that the classifier is most uncertain about for manual verification.

Additionally, we also compare the proposed risk-based solution (denoted by RISK) with the active learning solution (denoted by ACTL) on the achieved resolution quality provided with the same amount of human cost budget. Note that the ACTL solution would tune classifier parameters after additional manual verification, thus can potentially improve classification accuracy, while RISK would not.

We used the real datasets DBLP-Scholar [1] and Abt-Buy [2] in the empirical study. As usual, we use the standard blocking technique to filter the instance pairs unlikely to match. After blocking, the DBLP-Scholar workload contains totally 41416 instance pairs, and the Abt-Buy workload contains totally 20314 instance pairs. We employ SVM as the machine classifier. On DBLP-Scholar, we use the Jaccard similarity over the attributes *title* and *authors*, the edit distance over the attributes *title*, *authors* and *venue*, and the number equality over *publication year* as the input features for SVM. With only 1% of input data as training data, the achieved precision and recall of the SVM classifier are 0.917 and 0.875 respectively. On Abt-Buy, we use the Jaccard similarity and edit distance over the attributes *product name* and *description* respectively as the input features for SVM. With only 2% of input data as training data, the achieved precision and recall are 0.567 and 0.338 respectively. In the implementation of risk analysis, the confidence level θ is set to 0.8. Since a valid match probability should be between 0 and 1, we transform the inferred normal distribution to a *truncated normal distribution* in the range of 0 to 1 [15].

The comparative results on pick-up accuracy are presented in Figure 5.1. It can be observed that provided with the same amount of budget, CVAR consistently picks up more mislabeled pairs than BASE and UNCT. Since both BASE and UNCT reason about the risk based on the match expectation estimated by the machine, it should not be surprising that they perform similarly. The improvement margins of CVAR over the alternatives first enlarge with the increase of budget, but then gradually narrow down as expected. Since the number of mislabeled pairs decreases with additional manual inspections, the performance difference between different approaches tend to decrease as

---

[1] https://dbs.uni-leipzig.de/file/DBLP-Scholar.zip
[2] https://dbs.uni-leipzig.de/file/Abt-Buy.zip
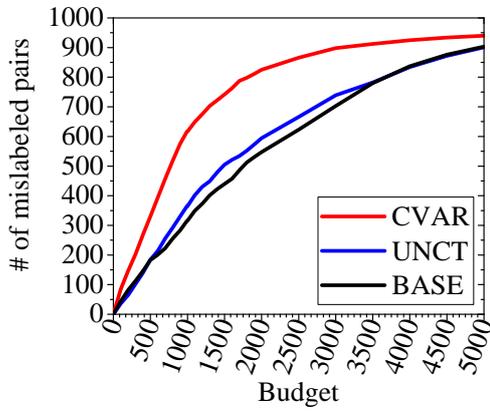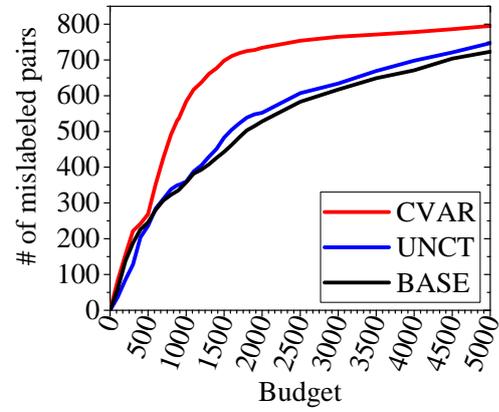
(a) The DBLP-Scholar dataset. (b) The Abt-Buy dataset.

Figure 5.1: Pick-up accuracy comparison.
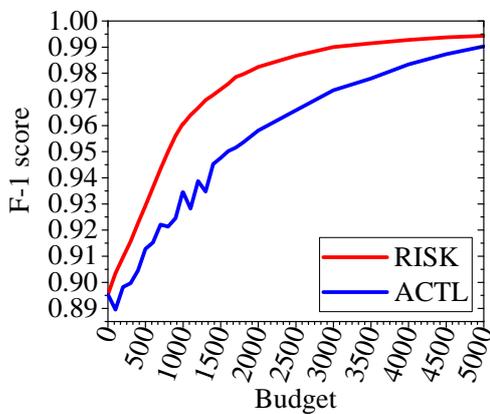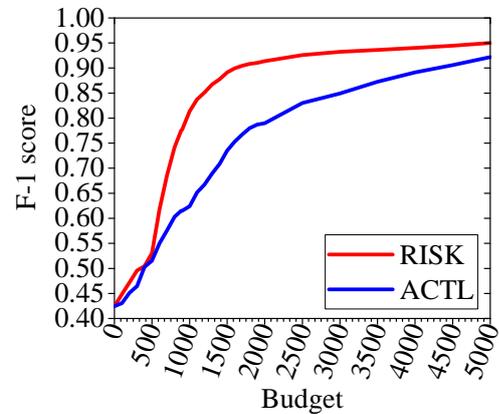
(a) The DBLP-Scholar dataset. (b) The Abt-Buy dataset.

Figure 5.2: Resolution quality comparison between RISK and ACTL.

well. These experimental results clearly validate the efficacy of the proposed risk model.

The comparative results on resolution quality, measured by the F-1 metric, between RISK and ACTL, are also presented in Figure 5.2. The achieved quality is measured on the results consisting of both manually labeled pairs and the pairs labeled by the classifier. It can be observed that after initial iterations, RISK achieves considerably better quality than ACTL. Even though ACTL uses the additional labeled data to update its classifier, the marginal benefit of additional training data points drops quickly with the increase of budget as expected. These experimental results show that the risk-based approach can be more effective than the active learning approach in improving resolution quality.

6 Conclusion

In this paper, we propose to investigate the problem of human and machine cooperation for ER from a risk perspective. We have presented a risk model and empirically validated its efficacy. It is worthy to point out that the proposed risk-based framework can be potentially generalized for other classification tasks. It is interesting to investigate its application in the scenarios besides ER in future work.

References

[1] P. Christen, *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer Science & Business Media, 2012, ch. 2, pp. 32–34.

[2] R. Singh, V. V. Meduri, A. Elmagarmid, S. Madden, P. Papotti, J.-A. Quiané-Ruiz, A. Solar-Lezama, and N. Tang, "Synthesizing entity matching rules by examples," *Proceedings of the VLDB Endowment*, vol. 11, no. 2, pp. 189–202, 2017.

[3] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1483–1494, 2012.

[4] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning," *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 269–278, 2002.

[5] A. Jain, A. D. Sarma, A. Parameswaran, and J. Widom, "Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace," *Proceedings of the VLDB Endowment*, vol. 10, no. 7, pp. 829–840, 2017.

[6] N. Vesdapunt, K. Bellare, and N. Dalvi, "Crowdsourcing algorithms for entity resolution," *Proceedings of the VLDB Endowment*, vol. 7, no. 12, pp. 1071–1082, 2014.

[7] C. Chai, G. Li, J. Li, D. Deng, and J. Feng, "Cost-effective crowdsourced entity resolution: A partial-order approach," *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pp. 969–984, 2016.

[8] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pp. 839–846, 2000.

[9] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden, "Scaling up crowdsourcing to very large datasets: a case for active learning," *Proceedings of the VLDB Endowment*, vol. 8, no. 2, pp. 125–136, 2014.

[10] J. O. Berger, *Statistical decision theory and Bayesian analysis.* Springer Series in Statistics, New York: Springer, 2nd ed., 1985.

[11] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[12] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of banking & finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[13] https://en.wikipedia.org/wiki/Normal_distribution.

[14] https://en.wikipedia.org/wiki/Inverse-gamma_distribution.

[15] J. Burkardt, "The truncated normal distribution," *Department of Scientific Computing Website, Florida State University*, 2014.