Bootstrapping Active Name Disambiguation with Crowdsourcing

Yu Cheng, Zhengzhang Chen, Jiang Wang, Ankit Agrawal, Alok Choudhary
EECS Department, Northwestern University
{ych133,zzc472,jwa368,ankit,choudhar}@eecs.northwestern.edu

ABSTRACT

Name disambiguation is a challenging and important problem in many domains, such as digital libraries, social media management and people search systems. Traditional methods, based on direct assignment using supervised machine learning techniques, seem to be the most effective, but their performances are highly dependent on the amount of training data, while large data annotation can be expensive and time-consuming requiring hours of manual inspection by a domain expert. To efficiently acquire labeled data, we propose a bootstrapping algorithm for the name disambiguation task based on active learning and crowdsourced labeling. We show that the proposed method can leverage the advantages of exploration and exploitation by combining two strategies, thereby improving the overall quality of the training data at minimal expense. The experimental results on two datasets DBLP and ArnetMiner demonstrate the superiority of our framework over existing methods.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Retrieval*

General Terms

Algorithms, Experimentation

Keywords

Name Disambiguation, Active Learning, Crowdsourcing, Bootstrapping

1. INTRODUCTION

Name disambiguation has been viewed as a very important problem in many domains, such as digital libraries, social media management [2] and people search systems. Given a large set of entity names, the task is to determine which names are referring to the same underlying entity. To solve this problem, there are many approaches proposed in the recent years [5, 8, 7, 6]. Generally existing methods mainly fall into three categories: supervised-based, unsupervised-based and constraint-based [9]. Among all the methods, the supervised-based approaches are considered to be the most effective ones [3]. However, their effectiveness are highly dependent on the amount of training data available.

For a large amount of training examples, the annotation work is expensive and time-consuming requiring hours of manual inspection by domain experts. In response, researchers have exploited active learning techniques to help with the labeling effort problems. In [3], Ferrera *et al.* proposed an active sampling strategy based on association rules to discriminate the author names. Wang *et al.* [9] introduced the active name disambiguation problem and presented the ADANA method to select the data for human labeling. By carefully selecting the informative samples, these approaches can largely reduce the amount of the data needed for constructing the training set. But the annotators still may get tired when asked to go through hundreds of hard-to-label samples in the labeling process, which is very tedious and error-prone. Moreover, previous active strategies for name disambiguation either select the data based on a single uncertainty measure [9] or just select the most potentially erroneous ones [3]. These criteria perform exploitation focuses on regions that are difficult to learn and would overlook those highly representative samples (exploration). This sampling bias may significantly limit the active learning performance.

On the other hand, social computing through services such as Amazon Mechanical Turk (MTurk) have made it possible for researchers to acquire sufficient quantities of crowdsourced labels. Crowdsourcing distributes problem solving to a broader community for requesting annotation and enable acquiring labeled data at less expensive cost.

In this paper, we propose a novel active learning algorithm, Active Data Augmentation, to exploit crowdsourcing techniques for name disambiguation. The proposed method combines discriminative features and crowdsourced labeling in a bootstrapping framework. Our hypothesis is that by combining the two strategies, we could balance the exploration and exploitation sampling. We first formulate the name disambiguation as a graph partitioning problem, and then we propose an algorithm to actively acquire labeled data by combining discriminative feature labeling and crowdsourcing with bootstrapping. Our approach is able to achieve both exploration and exploitation advantages, and

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507858 .

generate high quality training data at a minimal expense. The experimental results on two real-world datasets including DBLP and ArnettMiner demonstrate the effectiveness of our approach for the author name disambiguation task.

2. METHODOLOGY

In our approach, we first introduce a conditional pair-wise graph model for the name disambiguation problem. Then we use a graph partition based method to make the disambiguation decisions with some training samples. The training datasets are collected using our proposed active data augmentation method.

2.1 Conditional Name Disambiguation Model

We address the problem in the digital libraries domain. Given a person name, let $D = \{d_1, d_2, ..., d_n\}$ denote a collection of paper records which contain the name. By viewing each paper record d_i as a data point, it can be represented using l disambiguation features $\mathbf{x_i} = \{x_i^1, x_i^2, ..., x_i^l\}$ such as coauthors, paper titles, topics of articles, emails, affiliations etc. The objective of author name disambiguation is to group D into K clusters, where each cluster contains the references to a same author. Following the approaches proposed in [7, 6], we formalize the problem as a pairwise graph partition task. We generate a set of document feature pairs $X = \{(\mathbf{x_1}, \mathbf{x_2}), (\mathbf{x_2}, \mathbf{x_3}), ..., (\mathbf{x_{n-1}}, \mathbf{x_n})\}$, which is an observable variable. Let $y_{ij} \in Y$ be a hidden variable, representing whether two paper records refer to the same author, i.e. if $y_{ij} = 1$, then d_i and d_j belong to the same author, otherwise, they belong to different authors. The generated model would first learn maximum entropy for pairwise binary decisions, and then combine the information from the pairwise graph model using graph partitioning based methods so as to achieve a good global and consistent decision. The complete model for the conditional distribution of all binary match variables Y given all the observable data X can be expressed as:

$$P(Y|X) = \frac{1}{Z(\mathbf{x})} exp(\sum_{i,j,l} \lambda_l f_l(\mathbf{x_i}, \mathbf{x_j}, y_{ij}) + \sum_{i,j,k} \lambda_* f_*(y_{ij}, y_{jk}, y_{ik})) \quad (1)$$

where $Y = y_{ij} : \forall i, j$ and

$$Z(\mathbf{x}) = \sum_Y exp(\sum_{i,j,l} \lambda_l f_l(\mathbf{x_i}, \mathbf{x_j}, y_{ij}) + \sum_{i,j,k} \lambda_p f_p(y_{ij}, y_{jk}, y_{ik})) \quad (2)$$

$f_l(\mathbf{x_i}, \mathbf{x_j}, y_{ij})$ is a set of l feature function for the document pair (d_i, d_j). $f_p(y_{ij}, y_{jk}, y_{ik})$ is an equality transitivity function to ensure globally consistent configurations. The features for a pair of papers is described in Table 1.

2.1.1 Inference

Now, the problem is how to maximize the conditional probabilistic model (Eq. 1) with some training samples. Following [7], the inference of the model can be formulated as the problem of finding the graph partitions in which the nodes are the papers and the edge weights are the log clique potentials on the pair nodes $(\mathbf{x_i}, \mathbf{x_j})$ involved in their edges. Some methods, such as minimizing-disagreements correlations clustering [1, 10], Metropolis-Hastings (MH) [9] are introduced for the graph partitioning. We use the stochastic sampling to solve this problem, which is described in [6].

Table 1: Features for a pair of papers

Feature	Description
Title	similarity between titles of $\mathbf{x_i}$ and $\mathbf{x_j}$
Coauthors	indicator whether $\mathbf{x_i}$ and $\mathbf{x_j}$ share same author
Venue	whether $\mathbf{x_i}$ and $\mathbf{x_j}$ published in same venue
Affiliation	indicator whether $\mathbf{x_i}$ and $\mathbf{x_j}$ share same affiliation
NumCos	number of authors shared by $\mathbf{x_i}$ and $\mathbf{x_j}$

2.2 Active Learning with Crowdsourcing and Discriminative Feature Labeling

Given a bunch of unlabeled data, which samples should we select to query the user? In particular, in our problem, which document pairs (d_i, d_j) should be selected? And how can we get the labels from croudsourcing and discriminative feature labeling? The first problem can be addressed by using the active selection and the second is referred as acquiring labels based on two techniques.

2.2.1 Discriminative Features Labeling

The main purpose of discriminative features labeling is to identify some discriminative features which can help to determine whether the publications are written by the same author. Because we have not manually confirmed the labels, we refer to them as feature labels. Our idea is similar in spirit to the work proposed in [5], which intended to find "pure cluster" or "atomic cluster" in which publications must be correctly grouped (high precision) but might be further grouped in the process of clustering (possible low recall). For a pair of feature vectors ($\mathbf{x_i}$,$\mathbf{x_j}$), we define the features similarity $sim(\mathbf{x_i}, \mathbf{x_j})$ as:

$$sim(\mathbf{x_i}, \mathbf{x_j}) = \frac{\sum_m^M f_m(\mathbf{x_i}, \mathbf{x_j})}{M} \quad (3)$$

where $f_m(\mathbf{x_i}$,$\mathbf{x_j}) \in [0, 1]$, is the normalized score of m-th co-feature for the pair ($\mathbf{x_i}$,$\mathbf{x_j}$), and M is the number of co-features: Coauthors, Title, Venue, Affiliation and NumCoauthor. Table 1 summarizes this in details. In each step, the top-s pairs ($\mathbf{x_i}$,$\mathbf{x_j}$), with high similarity scores will be selected to labeled by feature labeling, i.e. $y_{ij} = 1$.

2.2.2 Crowdsouced Labeling

To acquire crowdsourced labels, we use the service from Amazon's Mechanical Turk. Users on the service received compensation ten cents for labeling a pair of papers with a list of paper information (title, authors, author affiliation, email, venue, and year). For each of these pairs, we ask five different users from AMT to label it as yes or no, which indicates whether the two authors are the same or not. Furthermore, we require that each crowdsourced training samples at least receive the same label by two users, thereby providing more certainty in the acquired label. This acts as a simple quality control for filtering out bad data from disinterested or exploitative users.

2.2.3 Uncertainty Measure

A straightforward solution for active selection is to select the most uncertain document pairs. According to Eq. 4, we could have the probability of two documents belonging to

the same cluster, i.e.,

$$p(y_{ij} = 1|\mathbf{x_i}, \mathbf{x_j}) = \frac{1}{Z_l} exp(\sum_l w_l f_l(\mathbf{x_i}, \mathbf{x_j}, y_{ij} = 1)) \quad (4)$$

If $p(y_{ij} = 1|\mathbf{x_i}, \mathbf{x_j}) = 0.5$, we say that the disambiguation model is the most uncertain about the document pair. $p(y_{ij} = 1|\mathbf{x_i}, \mathbf{x_j}) = 1$ and $p(y_{ij} = 1|\mathbf{x_i}, \mathbf{x_j}) = 0$, respectively, denote that the model is confident in that the two documents should be clustered together and should not be clustered. Based on the probability, we define an confidence score for a pair $(\mathbf{x_i}, \mathbf{x_j})$ as:

$$uncertainty(\mathbf{x_i}, \mathbf{x_j}) = |p(y_{ij} = 1|\mathbf{x_i}, \mathbf{x_j}) - 0.5| \quad (5)$$

An uncertain document pair will have a low confidence score in this case. Therefore, we select the k most uncertain document pairs with the lowest confidence scores according to Eq. 5.

2.2.4 Active Data Augmentation

Algorithm 1 illustrates our proposed technique with bootstrapping, which iteratively perform the training and evaluation to augment the training data. Given an unlabeled document pair set $\mathcal{U}$, let L denote the set of instances which return the top-s of $sim(\mathbf{x_i}, \mathbf{x_j})$. From this result, the top-s most confident predictions with discriminative feature labeling are added to L. We then update $\mathcal{U}$ by removing all instances also found in L, leaving only unlabeled examples in $\mathcal{U}$. The algorithm then iterates over the following steps. The top-k uncertain predictions are crowdsourced to obtain labels and then added into L. The algorithm terminates when the maximum number of iterations is reached or a given query budget is exhausted.

Algorithm 1 Active Data Augmentation Algorithm: Augments Training Data with Crowdsourcing

1: **Input**: Unlabeled set $\mathcal{U}$, parameters s and k.
2: **Output**: A selected dataset L from $\mathcal{U}$ with labels.
3: Initialize $L = \emptyset$;
4: Select the top-s confident instances and their labels L_f based on Eq. 3;
5: $L = L \bigcup L_f, \mathcal{U} = \mathcal{U} - L_f$.
6: Do
7: select top-k instances L_c with lowest confidence score based on Eq. 5;
8: query and get their crowdsourced labels;
9: $L = L \bigcup L_c, \mathcal{U} = \mathcal{U} - L_c$;
10: Until
11: the augment process is stopped.

3. EXPERIMENTS

In this section, we first describe the datasets and the baseline methods. Then, we compare the performance of our approach with the baseline methods. Finally we analyze the effectiveness of discriminative feature labeling.

3.1 Datasets

To evaluate our active data augmentation strategy, we use two collections of references derived from DBLP and ArnetMiner. These collections contain several ambiguous groups (groups of references with ambiguous author names). The first collection, hereafter referred to as DBLP, contains 4,287 references associated with 220 distinct authors. This means an average of approximately 20 papers per author. Its original version was created by Han *et al.* [4]. The second collection, hereafter referred to as ArnetMiner, was collected from the ArnetMiner.org [8] and labeled by more than 30 annotators. It contains 6,370 papers with 100 author names. Each paper is associated with a set of attributes: coauthor lists, title, publication venue, publication year, references, paper content, and affiliations.

Table 2: Average results on DBLP dataset with different sampling strategies

Method	Accuracy	Macro-F1
Random Selection	0.745	0.702
Random Selection, s=30	0.793	0.777
Active Associative Sampling	0.822	0.754
Active Selection	0.788	0.746
Active Data Augmentation, s=30	0.868	0.831

Table 3: Average results on ArnetMiner dataset with different sampling strategies

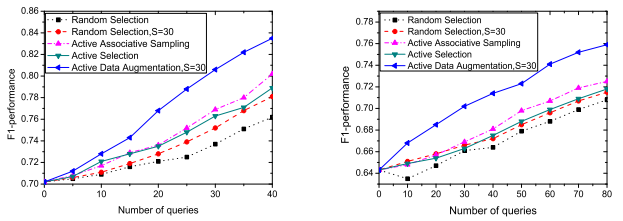
Method	Accuracy	Macro-F1
Random Selection	0.672	0.641
Random Selection, s=30	0.674	0.657
Active Associative Sampling	0.706	0.679
Active Selection	0.731	0.706
Active Data Augmentation, s=30	0.776	0.754

3.2 Baseline Methods

We compare the proposed approach with baseline results, with seed sets generated using discriminative feature labeling and labels from crowdsourcing. For each dataset, when evaluating the effectiveness of discriminative feature labeling, we use some number of initial instances from discriminative feature labeling. When using crowdsourced labels, we requested a total of 31,000 labels on 12,000 pairs of documents that were randomly chosen from each dataset, respectively. After employing the validation steps described in subsection 2.2.1, around 9,500 labels remained for each dataset. Our method selects the instances using the uncertainty strategy and uses 30 initial instances from feature labeling. We use four baseline methods in our experiments: (1) Random Selection: randomly select the instances to be labeled without initial seed data; (2) Random Selection, s=30: randomly select the instances to be labeled with 30 initial instances from discriminative feature labeling; (3) Active Associative Sampling: the method used in [3]; (4) Active Selection: select the instances using uncertainty measure without seed data. At each iteration, the training set is augmented with a number of k instances.

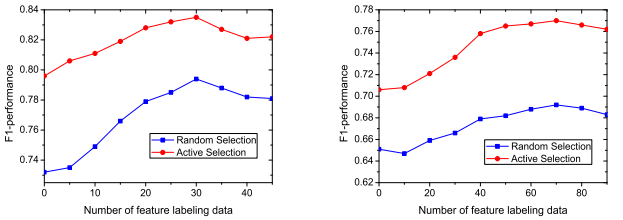
3.3 Results

In order to evaluate the effectiveness of our method, we use two evaluation metrics: accuracy and Macro-F1. Accuracy is the proportion of correctly disambiguated references. For Macro-F1, the performances are first calculated for individual author names and then averaged over all authors. In particular, we compare the results of the proposed algorithm with the baseline methods. Tables 2 and 3 show the results of name disambiguation on the two datasets af-



(a) DBLP (b) ArnetMiner

Figure 1: Performance comparison on DBLP and ArnetMiner datasets.



(a) DBLP (b) ArnetMiner

Figure 2: Effect of S on DBLP and ArnetMiner datasets.

ter 30% of training data are labeled. The proposed approach achieves an increase of at least 4.6% in accuracy and 3.9% in Macro-F1. Figures 1a and 1b show the variation of Macro-F1 score with the number of queries. We make the following observations: (1) starting with some seed data, the proposed Active Data Augmentation approach outperforms all baseline methods; (2) with some initial seed data, the random selection methods improve the performance but still results in increased error; (3) the method with single uncertainty strategy does not perform well, even comparing with the random selection with seed labels; (4) the active associative sampling cannot achieve much improvement. This indicates that a selection strategy using only discriminative feature labeling is insufficient and a strategy by considering both local and global information is necessary. Likewise, using only crowdsourced labels could not improve the results of active selection either, but combining discriminative feature labeling with crowdsourcing in our approach improves the results significantly.

3.4 Effect of S

We now study how the choice of the parameter s affects the performance. With fixed random sampling and active learning settings, we vary s, and observe the changes in the F-1 scores after 30% of training data are labeled. As shown in Figure 2a, as s increases from 0 to 30, the F-1 performance improves in both random and active settings. Similar trends can be observed in Figure 2b with a peak value 70. This implies that exploiting feature labeling to analyze the reviewer expertise can improve the classification accuracy. However, after it past a threshold, the accuracy tends to decrease. This might be due to over-fitting the training data with too much discriminative feature labeling.

4. CONCLUSION

We have present an unified framework for active name disambiguation by combining crowdsourcing and discriminative feature labeling. This bootstrapping method balances the exploration and exploitation advantages of both techniques, thereby improving the overall quality of the training data at minimal expense. Experimental results on two different genres of data sets showed that our proposed method outperforms the baseline methods.

5. REFERENCES

[1] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *Proceedings of the 43th Symposium on Foundations of Computer Science*, FOCS '02, pages 238–, Washington, DC, USA, 2002. IEEE Computer Society.

[2] Y. Cheng, K. Zhang, Y. Xie, A. Agrawal, W.-k. Liao, and A. Choudhary. Learning to group web text incorporating prior information. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 212–219. IEEE Computer Society, 2011.

[3] A. A. Ferreira, R. Silva, M. A. Gonçalves, A. Veloso, and A. H. Laender. Active associative sampling for author name disambiguation. In *Proceedings of the 12th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '12, pages 175–184, New York, NY, USA, 2012. ACM.

[4] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries*, pages 296–305, 2004.

[5] H. Han, H. Zha, and C. L. Giles. Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries*, pages 334–343. ACM, 2005.

[6] P. Kanani, A. McCallum, and C. Pal. Improving author coreference by resource-bounded information gathering from the web. In *Proceedings of the 20th International Joint Conference on Artifical intelligence*, IJCAI'07, pages 429–434, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[7] A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to noun coreference. In *Advances in Neural Information Processing Systems 17*, pages 905–912. MIT Press, 2005.

[8] J. Tang, A. Fong, B. Wang, and J. Zhang. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):975–987, 2012.

[9] X. Wang, J. Tang, H. Cheng, and P. S. Yu. Adana: Active name disambiguation. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pages 794–803, Washington, DC, USA, 2011. IEEE Computer Society.

[10] Y. Yang, J. Wang, and A. E. Motter. Network observability transitions. *Phys. Rev. Lett.*, 109:258701, Dec 2012.