Active Learning from Crowd in Document Screening

Evgeny Krivosheev
University of Trento, Italy
evgeny.krivosheev@unitn.it

Burcu Sayin
University of Trento, Italy
burcu.sayin@unitn.it

Alessandro Bozzon
TU Delft, The Netherlands
a.bozzon@tudelft.nl

Zoltán Szlávik
myTomorrows, The Netherlands
zoltan.szlavik@mytomorrows.com

Abstract

In this paper, we explore how to efficiently combine crowdsourcing and machine intelligence for the problem of document screening, where we need to screen documents with a set of machine-learning filters. Specifically, we focus on building a set of machine learning classifiers that evaluate documents, and then screen them efficiently. It is a challenging task since the budget is limited and there are countless number of ways to spend the given budget on the problem. We propose a multi-label active learning screening specific sampling technique -*objective-aware sampling*- for querying unlabelled documents for annotating. Our algorithm takes a decision on which machine filter need more training data and how to choose unlabeled items to annotate in order to minimize the risk of *overall classification* errors rather than minimizing a single filter error. We demonstrate that objective-aware sampling significantly outperforms the state of the art active learning sampling strategies.

1 Introduction

In this paper, we tackle the problem of screening a finite pool of documents, where the aim is to retrieve relevant documents *satisfying a given set of predicates* that can be verified by human or machines (Fig. 1). In this context, if a document does not satisfy at least one predicate, it is treated to be irrelevant. A predicate represents a property, a unit of meaning, given in natural language (e.g., "find papers that measure cognitive decline"). By this means a predicate might be interpreted in a variety of ways in text, so making keywords-based search hard to reach high recall while keeping a decent level of precision [42]. We interpret the screening problem as high recall problem, i.e., the aim is to retrieve all relevant documents maximizing precision. The screening finds application in many domains, such as i) in systematic literature reviews [36]; ii) database querying - where items filtered in/out based on predicates [25]; iii) hotel search - where the hotels retrieve based upon filters of interest [18]. Consequently, the document screening is an instance of finite pool binary classification problems [23], where we need to classify a finite set of objects minimizing cost. A bottleneck of the screening process is the predicate evaluation, i.e., identifying which of the given predicates are satisfied in a current document. For example, in literature reviews, authors validate predicates, however, this is time-consuming, exhaustive, and very expensive [12, 15, 17, 32].

An effective technique to solve screening problems is *crowdsourcing* where the crowd can solve even complex screening tasks with high accuracy and lower cost compared to expert screening [3, 15, 17, 22, 25, 26, 30, 34, 36]. However, achieving a good performance in crowd-based screening requires a deep understanding of how to design tasks and model their complexity [8, 39, 27, 40], how

NeurIPS 2020 Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation.

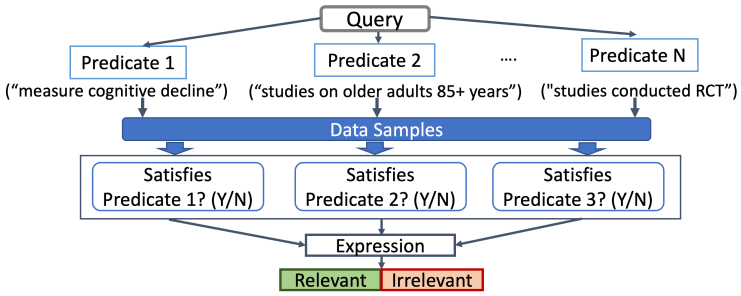

Figure 1: Predicate-based screening process of documents

to test and filter workers [2], how to aggregate results into a classification decision [6, 7, 9, 14, 17, 19, 20, 21, 24, 37, 43], or how to improve worker engagement [10, 11, 28].

Machine learning (ML) algorithms have also made a very impressive progress in solving complex screening tasks. However, obtaining a sufficiently large set of training data is still a key bottleneck for accurate ML classifiers. Active learning (AL) [5] accelerates this process by minimizing the size of training data that is required to train better classifiers via selecting the most informative instances for annotation. The effectiveness of AL have been proven in many domains (see the surveys [1, 33]), but most of the work considers single-label cases while multi-label AL problems have been far less investigated. The challenge in applying AL to a multi-label classification problem is that the algorithm should measure the unified informativeness of each unlabeled item across all labels. The state of the art multi-label AL strategies follow an object-wise (global) labeling, where the AL algorithm first finds the relevance scores (i.e. confidence-based or disagreement-based scores) of $< item, label >$ pairs, and then aggregates these scores to find the informativeness of items [4, 13, 35, 38, 41]. However, it may ignore the interaction between labels [31].

Original contribution. We investigate how to efficiently combine crowdsourcing and ML for item screening. It is a challenging task since the budget is limited and there are countless number of ways to spend it on the problem. We propose a multi-label AL screening specific sampling technique for querying unlabelled items for annotating. Our algorithm takes a decision how to choose unlabeled data to annotate by crowd workers in order to maximize the performance of a screening task. Unlike existing multi-label AL approaches that rely on global labeling, we choose the local labeling method, where for each label (predicate in our case) we determine the relevancy to each item.

2 Approach

2.1 Problem Statement

We model the problem as the input tuple (P, UI, B, E, AL), where $P = \{p_1, .., p_n\}$ is the set of inclusive filters expressed in natural language, e.g., *P = ("papers study older adults", "papers study mental disorders")*, UI is the set of unlabeled items (documents) that needed to be labeled as either "IN" or "OUT" of scope, B is an available budget for collecting training data for classifiers A (we do not have training data in the beginning), AL is an active learning strategy, and E is a conjunctive expression of predicates. The output is the tuple (LI, A), where LI is the labeled items, A is the set of trained binary ML classifiers that evaluate if an item describes the predicates. The screening operates on each unlabeled item estimating how much predicates P match the current document i. If at least one of the predicates does not match, the document is considered as irrelevant. We can compute the probability that a document i should be excluded: $Prob(i \in OUT) = 1 - \prod_{p \in P} Prob(i_p \in IN)$. In this work, we propose a new AL strategy for multi-predicate classification and evaluate it under different settings. We observe that rather than having one ML algorithm for the overall screening problem, it is often convenient to have a separate classifier A^p_{ml} for each predicate p. There are several reasons for this: besides being a requirement of some crowd-machine algorithms [23, 16] from budget optimisation to facilitating their reuse [4, 29].

2.2 Objective-Aware Sampling

We now discuss how ML classifiers for each predicate can be trained, given an ML algorithm (e.g., SVM). The most common AL method is *uncertainty sampling*, that selects items closest to a classifier's decision boundary [1]. We introduce here an *objective-aware sampling* technique particularly suited for screening problems and that, as we will show, can outperform uncertainty sampling. We motivate the need for – and the rationale of – *objective-aware sampling* with an example. We want to train two binary classifiers (A^1_{ml}, A^2_{ml}) to evaluate predicates p_1, p_2 respectively. As we iterate through the training set to learn the model, assume we reach the situation shown in Table 1. In AL fashion we then need to choose the next *(item, predicate)* to label.

Table 1: Objective-aware sampling queries items for a classifier based on overall screening uncertainty.

Item ID	$Prob(i_{p_1} \in IN)$ estimated by A^1_{ml}	$Prob(i_{p_2} \in IN)$ estimated by A^2_{ml}	True class (hidden)
1	0.99	0.98	IN
2	0.51	0.01	OUT
3	**0.51**	**0.99**	**OUT**
4	0.03	0.01	OUT

In this context, if we want to improve A^1_{ml} independently from the overall screening objective, then we adopt uncertainty sampling as per the literature, choosing either item 2 or 3. However, the classification errors of A^1_{ml} might have a bigger impact if predicate p_2 applies to the item (that is, if the item is IN scope according to p_2), because then the final decision rests on p_1 and therefore A^1_{ml}. In the table above, the risk of overall classification errors for item 2 is low as A^2_{ml} will filter out the item anyways, regardless of the opinion of A^1_{ml}. This is not the case for item 3 which is therefore a more promising case to pick for labeling. In other words, denoting with P^p the estimation by the ML classifier A^p_{ml} trained on predicate p, we choose the item for which the chance of making the *False Negative* error is the highest. In summary, we apply a combination of uncertainty and certainty sampling, but we look at the screening accuracy rather then only at the current classifier:

$$score(i, \bar{p}) = (1 - Prob(\hat{y}|i, A^{\bar{p}}_{ml})) \cdot \prod_{p \in \{P - \bar{p}\}} Prob(i_p \in IN) \tag{1}$$

where $\bar{p}$ is the predicate for which we want to improve the ML classification, $\hat{y} = argmax_y P(y|i)$ is the class label with the highest posterior probability , $\{P - \bar{p}\}$ the set of predicates except one we want to improve on. We then chose top-k (item, predicate) with the highest scores for labeling. Notice that objective-aware sampling applies when classifiers are trained for each predicate.

3 Experiments and Results

Methodology. We experiment objective-aware strategy with the goal of understanding whether there is an improvement in screening accuracy where we fix a budget that can be spent on crowd or expert annotations. Our baseline AL strategies include uncertainty sampling and random sampling.

To do this, we identify datasets with different characteristics in terms of accuracy that machine or crowd achieve, and in terms of predicate selectivity. The `Amazon Reviews` dataset [1] contains 5000 reviews on products, with information about the binary sentiment expressed in the reviews and whether a review was written on a book. In this screening task, we have two predicates: the `Books` predicate (with selectivity of 0.61) and the `Negative review` (with selectivity of 0.10), only 5% of reviews are therefore considered relevant. The rationale for this dataset is to have a scenario where the task is relatively easy for the crowd with the average crowd accuracy of 0.94. `Systematic Literature Review` (SLR) dataset [17] mimics the screening phase of the SLR process, where the abstracts returned from a keywords-based search need to be screened by researchers. This dataset contains 825 items for two predicates: `Does the experiment targeting and involving "older adults"?` (with selectivity of 0.58, crowd accuracy is 0.8) and `Does the paper describe an intervention study?` (with selectivity of 0.20, crowd accuracy is 0.6), and 17% of relevant items. Unlike the Amazon dataset, this task is hard for the crowd and

[1] https://github.com/TrentoCrowdAI/crowdsourced-datasets

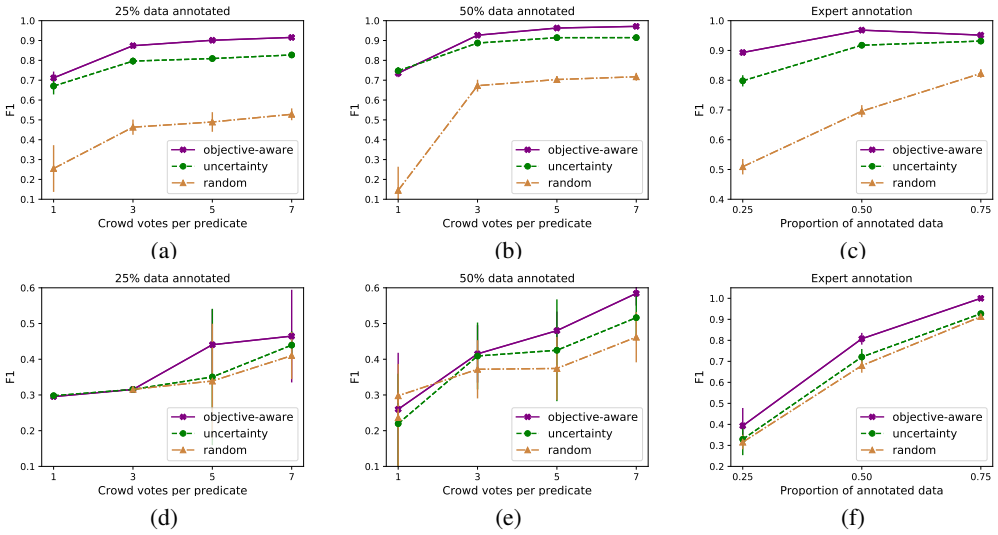

Figure 2: Results for Amazon (a, b, c) and SLR (d, e, f) datasets

even harder for ML classifiers. Each of the datasets was partially labeled by crowd workers (e.g., 1k reviews in the Amazon dataset) which allows us to estimate the accuracy distribution of workers on predicates and to simulate crowd annotations with the real answer probability distributions of the datasets and thus to analyze the proposed strategy under different conditions.

For each predicate, we trained an SVM classifier with TF-IDF features and “balanced" class weights. We use F1 score to measure the classification accuracy and report its average value over 10 runs. The source code and experiment results (including additional metrics) are available online [2]

Results. To analyze the proposed AL strategy, we run experiments over the available datasets varying the budget, the proportion of annotated data, and the quality of annotations. Figure 2 depicts the F1 screening score for 25% and 50% of annotated data samples. Along X-axis, we increase the number of crowd votes requested per (item, predicate) during the annotation process that followed by Majority Voting aggregation. Intuitively, the more crowd votes were queried per item the better accuracy of labels, and the more budget we need to spend considering a fixed proportion of annotated data. We observe for the Amazon dataset, where crowd accuracy is high, the objective-aware sampling always outperforms the baselines. Notably, with more accurate annotations (more crowd votes per predicate) our approach makes the accuracy gap between baselines even bigger thus outperforming uncertainty sampling by 7.8 and 8.8 points in F1 when collecting 3 and 7 votes/predicate respectively. SLR dataset is very difficult for both crowd and machines, and it is crucial to obtain high quality annotations. With less than 3 votes per predicate, the labels become highly inaccurate leading to poor screening accuracy. When we request 3 or more votes per predicate the results become much better and, as in the Amazon dataset, our objective-aware sampling strategy superiors both uncertainty and random sampling techniques by a big margin (Figure 2 d, c).

We further examine how objective-aware sampling behaves in case of noise-free annotations obtained by experts. Figure 2 c,f shows the resulting F1 scores for different proportions of labeled data. We can see that our approach outperforms the baselines for both Amazon and SLR datasets on all the settings. Thus for 50% of annotated data objective-aware sampling beats uncertainty sampling by 5.1 and 8.7 in F1 points for Amazon and SLR datasets respectively.

4 Conclusion

In this paper, we proposed and evaluated the objective-aware active learning strategy designed for screening classification and selecting efficiently item, predicate for annotating based on the overall classification objective. We demonstrated that objective-aware sampling outperforms uncertainty and

[2] https://github.com/Evgeneus/Active-Hybrid-Classificatoin_MultiPredicate

random AL techniques under different conditions. We further aim to examine more screening datasets, extend this study to other classes of screening problems and hybrid crowd-machine algorithms.

References

[1] Charu C. Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and Philip S. Yu. *Active Learning: A Survey, Data Classification: Algorithms and Applications*. CRC Press, 2014.

[2] Jonathan Bragg, Mausam, and Daniel S. Weld. Optimal testing for crowd workers. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '16, pages 966–974, Richland, SC, 2016.

[3] William Callaghan, Joslin Goh, Michael Mohareb, Andrew Lim, and Edith Law. Mechanicalheart: A human-machine framework for the classification of phonocardiograms. *Proc. ACM Hum. Comput. Interact.*, 2(CSCW):28:1–28:17, 2018.

[4] Everton Alvares Cherman, Grigorios Tsoumakas, and Maria-Carolina Monard. Active learning algorithms for multi-label data. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 267–279. Springer, 2016.

[5] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *J. Artif. Int. Res.*, 4(1):129–145, 1996.

[6] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C Applied Statistics*, 28(1), 1979.

[7] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Data fusion : Resolving conflicts from multiple sources. In *Procs of WAIM2013*. Springer, 2013.

[8] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 5–14, New York, NY, USA, 2017.

[9] Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. Crowd worker strategies in relevance judgment tasks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 241–249, New York, NY, USA, 2020.

[10] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. All those wasted hours: On task abandonment in crowdsourcing. WSDM '19, page 321–329, New York, NY, USA, 2019.

[11] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 10 2019.

[12] Julian PT Higgins and Sally Green. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

[13] Sheng-Jun Huang, Songcan Chen, and Zhi-Hua Zhou. Multi-label active learning: Query type matters. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 946–952. AAAI Press, 2015.

[14] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011.

[15] Evgeny Krivosheev, Valentina Caforio, Boualem Benatallah, and Fabio Casati. Crowdsourcing paper screening in systematic literature reviews. In *Procs of Hcomp2017*. AAAI, 2017.

[16] Evgeny Krivosheev, Fabio Casati, Marcos Baez, and Boualem Benatallah. Combining crowd and machines for multi-predicate item screening. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018.

[17] Evgeny Krivosheev, Fabio Casati, and Boualem Benatallah. Crowd-based multi-predicate screening of papers in literature reviews. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 55–64, 2018.

[18] Doren Lan, Katherine Reed, Austin Shin, and Beth Trushkowsky. Dynamic filter: Adaptive query processing with the crowd. In *Procs of Hcomp2017*. AAAI, 2017.

[19] Hongwei Li, Bin Yu, and Dengyong Zhou. Error rate analysis of labeling by crowdsourcing. In *Procs of ICML2013*, 2013.

[20] Chao Liu and Yi Min Wang. Truelabel + confusions: A spectrum of probabilistic models in analyzing multiple ratings. In *Procs of ICML2012*. ICML, 2012.

[21] Qiang Liu, Alexander T Ihler, and Mark Steyvers. Scoring workers in crowdsourcing: How many control questions are enough? In *Advances in Neural Information Processing Systems*, pages 1914–1922, 2013.

[22] Michael L Mortensen, Gaelen P Adam, Thomas A Trikalinos, Tim Kraska, and Byron C Wallace. An exploration of crowdsourcing citation screening for systematic reviews. *Research Synthesis Methods*, 2016. RSM-02-2016-0006.R4.

[23] An T Nguyen, Byron C Wallace, and Matthew Lease. Combining Crowd and Expert Labels using Decision Theoretic Active Learning. *Proceedings of the 3rd AAAI Conference on Human Computation (HCOMP)*, pages 120–129, 2015.

[24] Jungseul Ok, Sewoong Oh, Jinwoo Shin, and Yung Yi. Optimality of belief propagation for crowdsourced classification. In *Procs of ICML2016*, 2016.

[25] Aditya Parameswaran, Stephen Boyd, Hector Garcia-Molina, Ashish Gupta, Neoklis Polyzotis, and Jennifer Widom. Optimal crowd-powered rating and filtering algorithms. *Proc. VLDB Endow.*, 7(9):685–696, May 2014.

[26] Aditya G. Parameswaran, Hector Garcia-Molina, Hyunjung Park, Neoklis Polyzotis, Aditya Ramesh, and Jennifer Widom. Crowdscreen: Algorithms for filtering data with humans. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, page 361–372, New York, NY, USA, 2012.

[27] Rehab Qarout, Alessandro Checco, and Kalina Bontcheva. Investigating stability and reliability of crowdsourcing output. In *CEUR Workshop Proceedings*, 07 2018.

[28] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020.

[29] Jorge Ramirez, Evgeny Krivosheev, Marcos Baez, Fabio Casati, and Boualem Benatallah. Crowdrev: a platform for crowd-based screening of literature reviews. *arXiv preprint arXiv:1805.12376*, 2018.

[30] Benjamin L. Ranard, Yoonhee P. Ha, Zachary F. Meisel, David A. Asch, Shawndra S. Hill, Lance B. Becker, Anne K. Seymour, and Raina M. Merchant. Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, 29(1), 2014.

[31] Oscar Reyes, Carlos Morell, and Sebastin Ventura. Effective active learning strategy for multi-label learning. *Neurocomput.*, 273(C):494–508, January 2018.

[32] Margaret Sampson, Kaveh G Shojania, Chantelle Garritty, Tanya Horsley, Mary Ocampo, and David Moher. Systematic reviews can be produced and published faster. *Journal of clinical epidemiology*, 61(6):531–536, 2008.

[33] Burr Settles. Active learning literature survey. Technical report, 2010.

[34] Yalin Sun, Pengxiang Cheng, Shengwei Wang, Iain James Marshall, Hao Lyu, Matthew Lease, and Byron C Wallace. Crowdsourcing information extraction for biomedical systematic reviews. *4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2016.

[35] Deepak Vasisht, Andreas Damianou, Manik Varma, and Ashish Kapoor. Active learning for sparse bayesian multilabel classification. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 472–481, New York, NY, USA, 2014.

[36] Byron C Wallace, Anna Noel-Storr, Iain J Marshall, Aaron M Cohen, Neil R Smalheiser, and James Thomas. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association*, 24(6):1165–1168, 05 2017.

[37] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. pages 2035–2043, 2009.

[38] Jian Wu, Victor Sheng, Jing Zhang, Pengpeng Zhao, and Zhiming Cui. Multi-label active learning for image classification. 10 2014.

[39] Meng-Han Wu and Alexander J. Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *HCOMP*, 2017.

[40] Jie Yang, Judith Redi, Gianluca DeMartini, and Alessandro Bozzon. Modeling task complexity in crowdsourcing. In *Proceedings of The Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016)*, pages 249–258. AAAI, 2016.

[41] Chen Ye, Jian Wu, Victor Sheng, Pengpeng Zhao, and Zhiming Cui. Multi-label active learning with label correlation for image classification. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3437–3441, 09 2015.

[42] Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. Effective user interaction for high-recall retrieval: Less is more. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 187–196, New York, NY, USA, 2018.

[43] Dengyong Zhou, John C. Platt, Sumit Basu, and Yi Mao. Learning from the wisdom of crowds by minimax entropy. In *Procs of Nips 2012*, 2012.