WHEN IN DOUBT ASK THE CROWD

EMPLOYING CROWDSOURCING FOR ACTIVE LEARNING

Mihai Georgescu, Dang Duc Pham, Claudiu S Firan, Ujwal Gadiraju, Wolfgang Nejdl

L3S Research Center
Leibniz Universität Hannover

Free Search
Toward Computer Science Ideas

Introduction

- Integrated framework for active learning using crowd assigned labels, gathered on demand as training data for an automatic method
- Enable an automatic method and human labelers to work together towards improving their performance

- Identify the major challenges that can arise when deploying such a framework

- Provide extensive experiments using various automatic methods that learn to perform a task by exploiting the wisdom of the crowds

Crowdsourcing

- **Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.[1]**

- **The crowd workers are motivated by a small financial incentive**
- **Usually done via microtask platforms such as Amazon's Mechanical Turk or Crowdflower**
- **Requester posts HITs that are solved by workers for a financial reward**
- **Unknown workers with various expertise can replace domain experts**
- **Advantages: cost effective, workers availability and diversity**
- **Disadvantages: questionable quality of work**

1) crowdsourcing.typepad.com

Automatic Labels vs. Crowd Labels

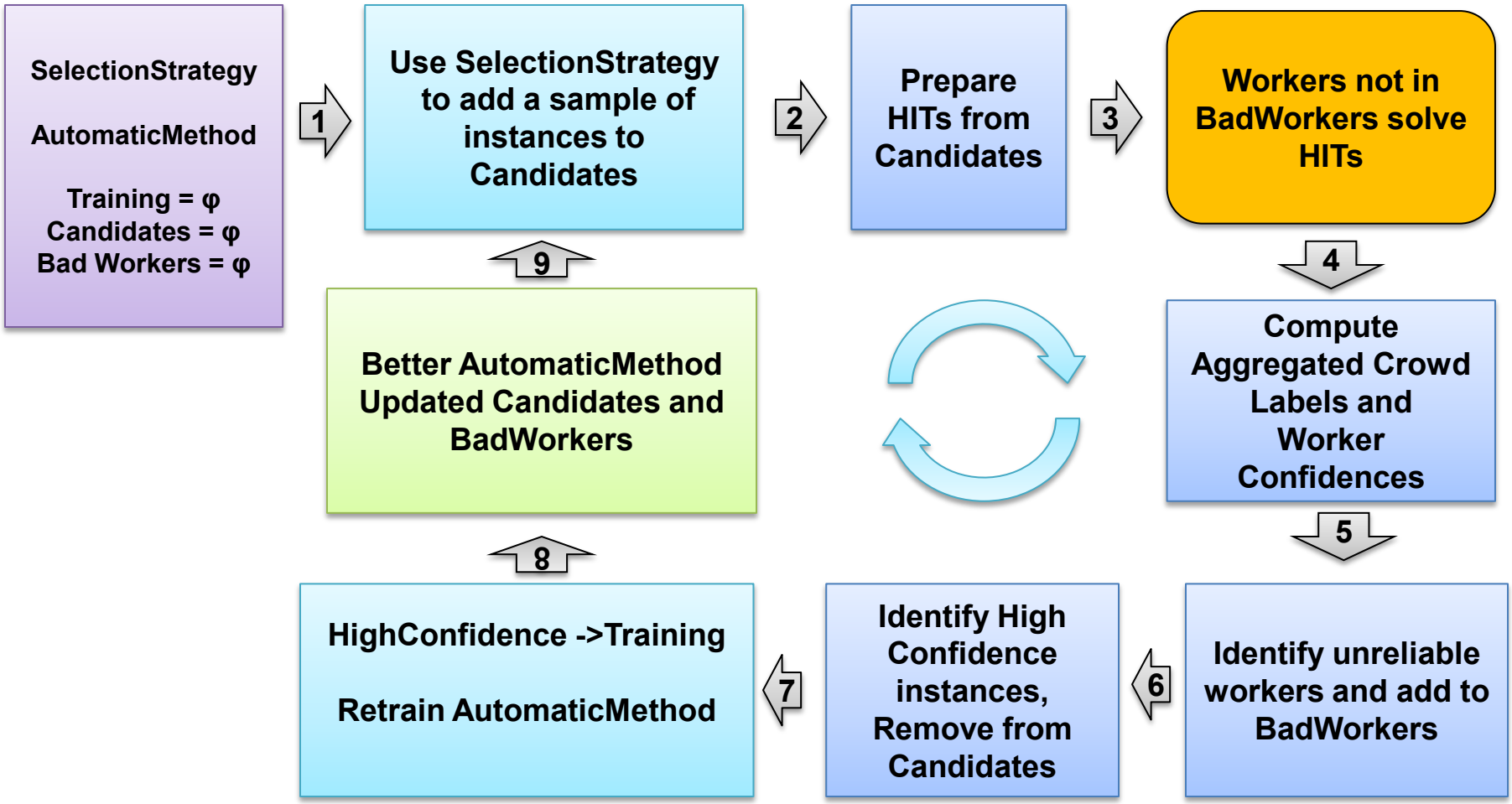
- Gather labels from the crowd in an active learning manner, for training an automatic method
- For each instance gather multiple labels and aggregate them

- Crowd Labels (CL) - Binary labels, aggregation of labels from all workers
- Crowd Soft Labels (CSL)
  - [0,1] value for the confidence we have in the crowd label
  - Can incorporate the notion of worker confidence (reliability)

- Automatic Labels(AL) - Binary label produced by the Automatic Method
- Automatic Soft Labels(ASL)
  - [0,1] value for the confidence of the automatic label
  - Used by aSelection Strategy for finding instances for which labels are needed to improve the Automatic Method

- Goal: Have AL as close as possible to CL

Continuous Active Learning Process



Specific task: deduplication of scientific publications

- Objective: Automatic deduplication of scientific publications
- Solution: use the proposed method and let an automatic algorithm actively learn from the crowd how to deduplicate

- Considered instances: pairs of publications described by metadata
- List of fields a publication might have: Title, Subtitle, By, In, Type, Publisher, Organization, Abstract
- Labels: a pair contains duplicate or not-duplicate publications

Automatic Methods

- Duplicates Scorer
  - Produces an ASL based on an epsilon-adjusted mean of field similarities
  - Using as parameters the weights of the fields
  - Final assignment comes from comparing the ASL to a threshold

- Classifiers:
  - ASL is the classifier confidence in the class assignment
  - Naïve Bayes, Decision Tree or SVM
  - Features: Similarities between fields (Needleman-Wunch or Jaccard)
  - Each instance (pair of publications) has 8 features

Learning from the crowd

- Automatic method provides an ASL, indicating confidence
- Use ASL to select instances according to a Selection Strategy
- Use all reliable labels (HighConfidence) to re-train

- DuplicatesScorer
  - Start with a common sense parameter choice
  - In each round when re-training, take into consideration the parameters learned in the previous round and used for the selection
- Classifiers
  - Start with a random sample
  - Re-training uses all the reliable acquired labels

Evaluation

- Dataset
- Inter-Agreement of labelers
- Performance of different Automatic Methods
- Resource Allocation per Active Learning Round
- Selection Strategy

Dataset

Pairs of publications from different data sources: DBLP, CiteSeer, BibSonomy,TibKat as in the Freesearch system (dblp.kbs.uni-hannover.de)

Ground Truth:

- 363 pairs labeled by 3 experts: 101 dupl, 262 non-dupl

Crowd Data

- includes ground truth
- 2070 pairs with at least 3 crowd labels
- 570 pairs with 7 crowd labels
- MV : 804 dupl, 1264 non-dupl

Mechanical Turk Task

amazon mechanical turk™
Artificial Artificial Intelligence

[Show Diff] [Full Text]
Title: Comparing Heuristic, Evolutionary and Local Search Approaches to Scheduling

Authors: Soraya Rana, Adele E. Howe, L. Darrell, Whitley Keith Mathias
Venue: Proceedings of the Third International Conference on Artificial Intelligence Planning Systems, Menlo Park, CA
Publisher: The AAAI Press
Year: 1996
Language: English
Type: conference

Abstract: The choice of search algorithm can play a vital role in the success of a scheduling application. In this paper, we investigate the contribution of search algorithms in solving a real-world warehouse scheduling problem. We compare performance of three types of scheduling algorithms: heuristic, genetic algorithms and local search.

[Show Diff]
Title: Comparing Heuristic, Evolutionary and Local Search Approaches to Scheduling.

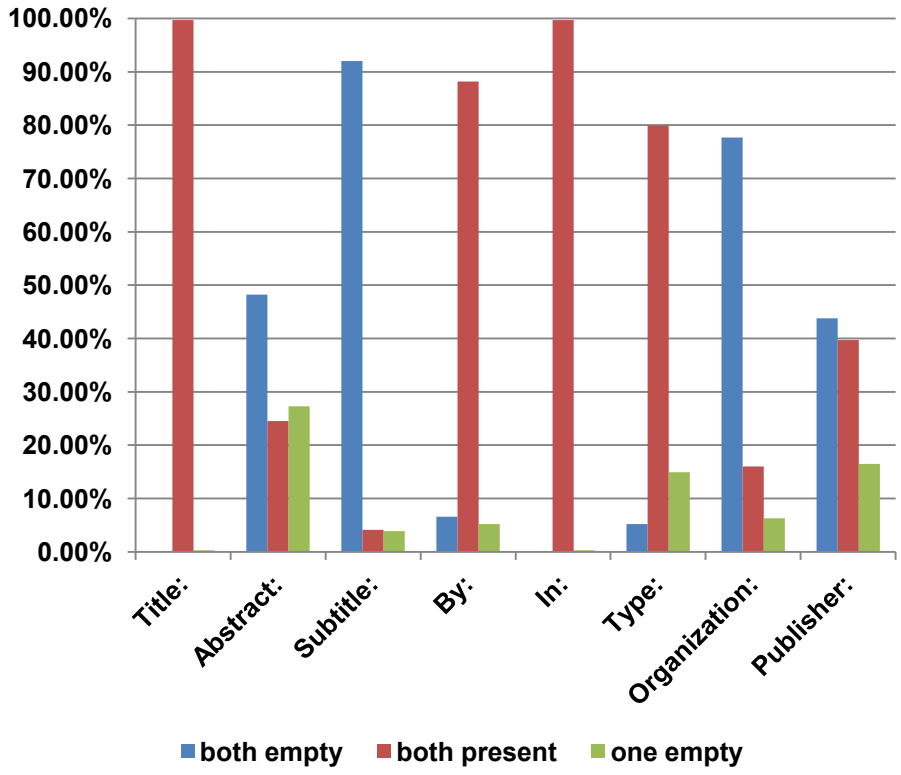
Authors: Soraya B. Rana, Adele E. Howe, L. Darrell Whitley, Keith E. Mathias
Book: AIPS Pg. 174-181 [Contents]
Year: 1996
Language: English
Type: conference (inproceedings)

After carefully reviewing the publications metadata presented to you, how would you classify the 2 publications referred:

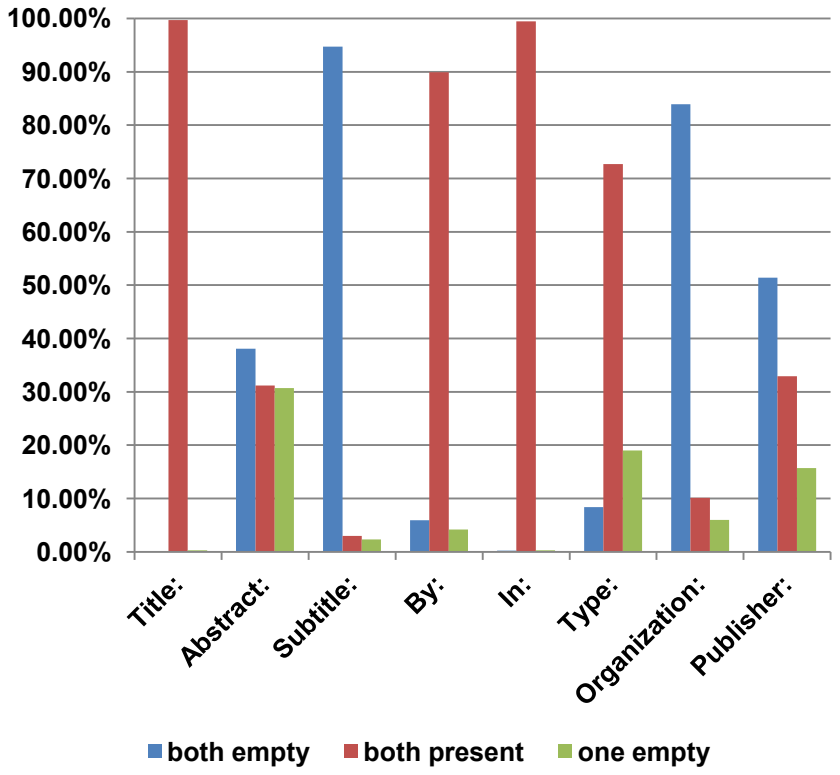
Judgment for publications pair:
○ Duplicates
○ Not Duplicates

Dataset Statistics

Ground Truth Field Distribution



Crowd Data Field Distribution

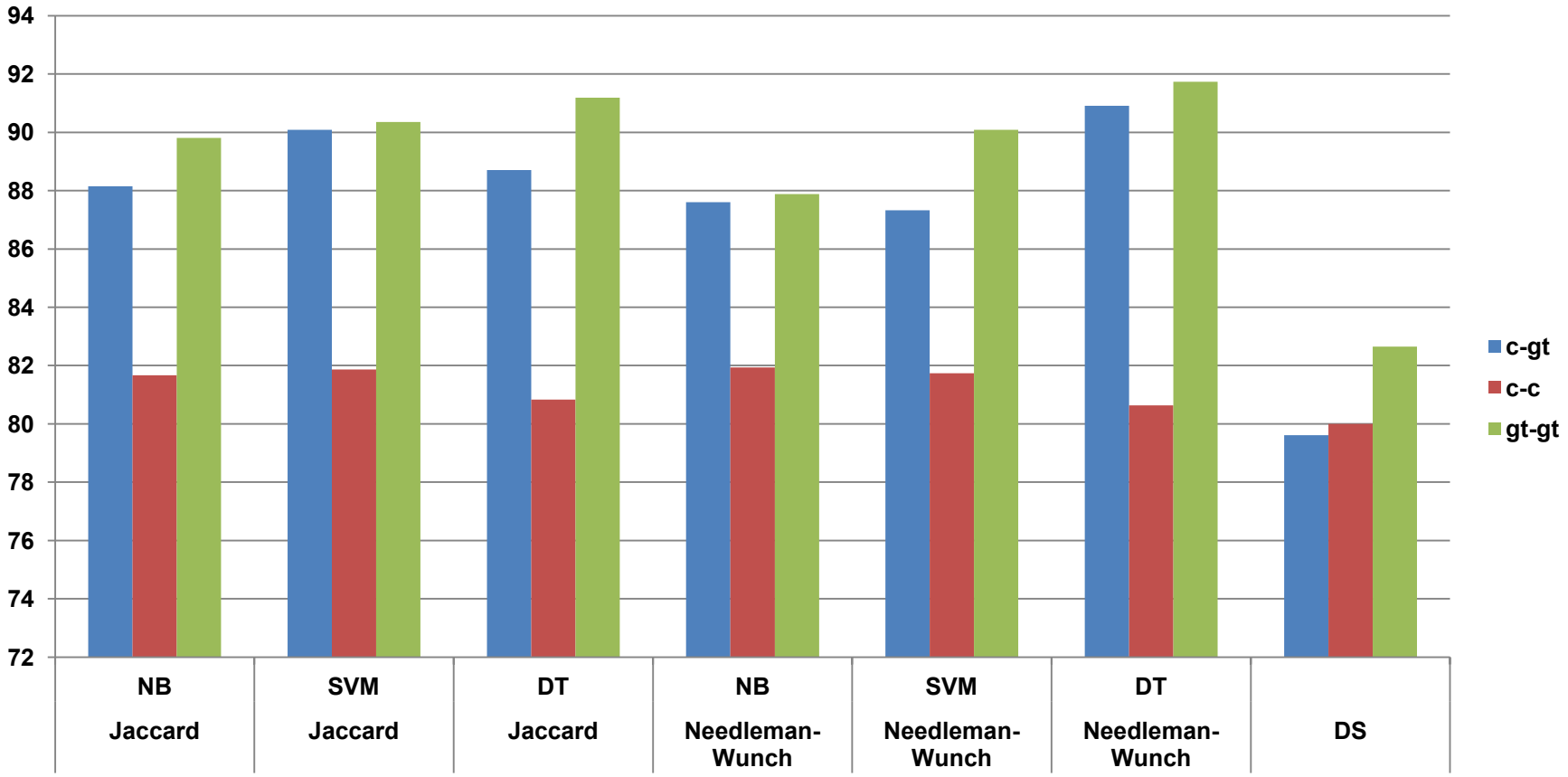


Agreement

Labels	Instances	Fleiss Kappa	Krippendorf Alpha
Experts on ground Truth			
3	301	0.827	0.827
Crowd on Ground Truth			
3	358	0.526	0.526
4	358	0.526	0.526
5	358	0.503	0.511
6	337	0.478	0.499
7	285	0.47	0.492
Crowd on Training Data			
3	2064	0.282	0.282
4	560	0.506	0.303
5	560	0.499	0.319
6	528	0.495	0.331
7	425	0.477	0.338

- **Experts are more in agreement than crowd workers**
- **On the ground truth more than 3 crowd workers leads to less agreement**
- **On the larger crowd data, 5 workers are better agreeing than 3, but in less agreement than 7**
- **There is a limit after which introducing more workers is detrimental to the agreement**

Accuracy of different methods



Attribute Selection

	Leave-1-out	Chi-squared	Info gain
title	0.73705	671.5102	0.35174
abstract	0.79656	156.4479	0.07633
subtitle	0.7905	0	0
by	0.78223	163.7297	0.08084
in	0.78815	89.1981	0.04172
type	0.78113	0	0
organization	0.79284	2.665	0.00124
publisher	0.79256	29.0746	0.01355

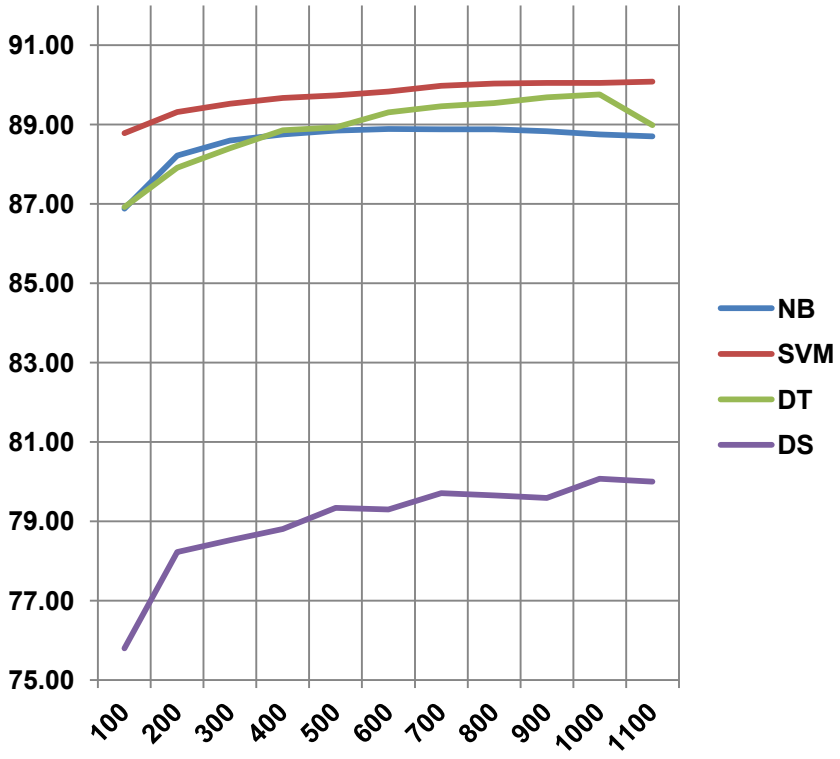
Best fields:

- **Title, by, abstract, in for classifiers**
- **Title, by, type, in for DS**

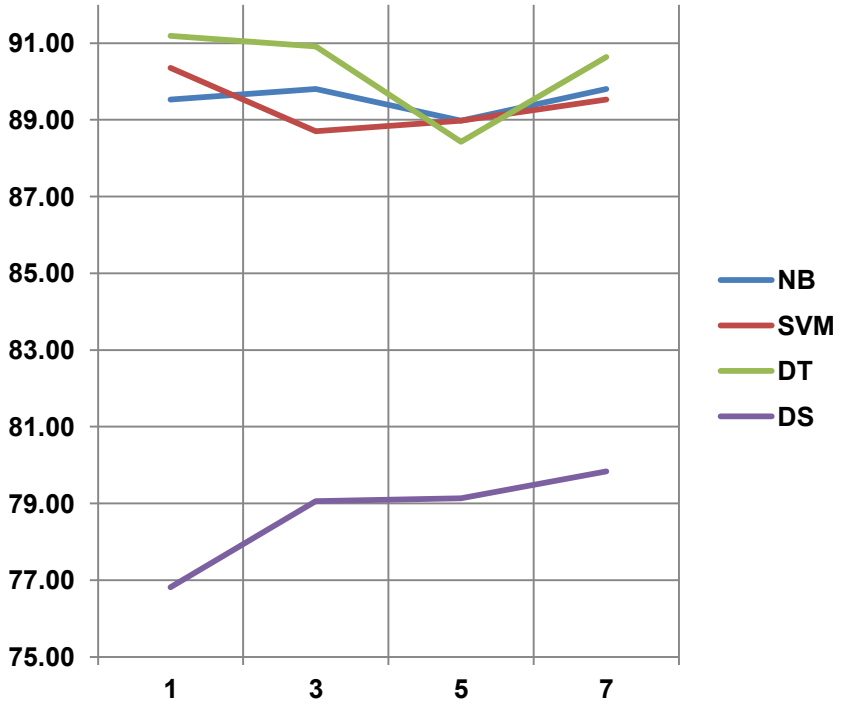
Matches the field distribution
Fields for which values are present in both publications are more important

Resource Allocation

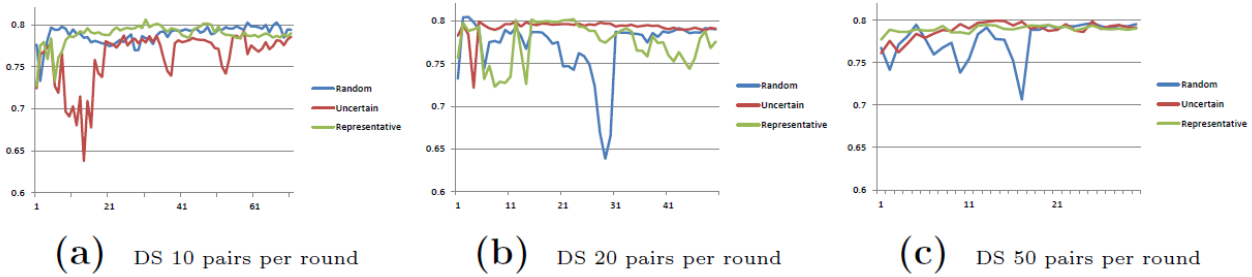
Number of tasks per round



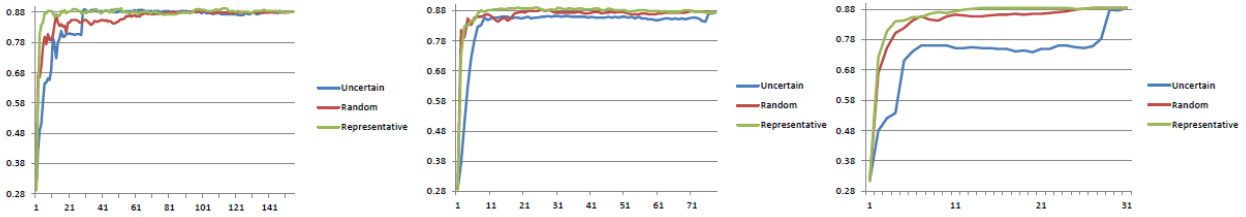
Number of assignments per task



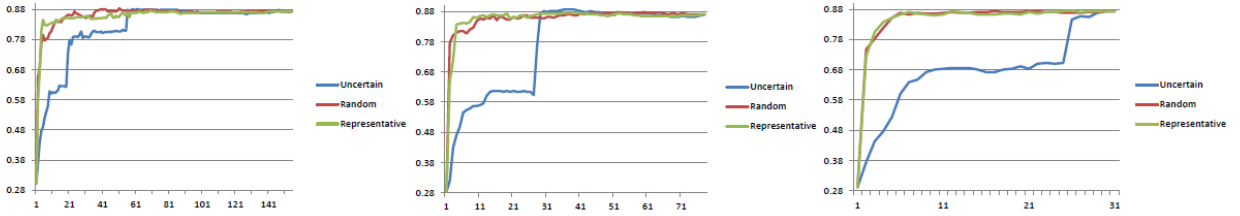
Selection Strategies



(a) DS 10 pairs per round (b) DS 20 pairs per round (c) DS 50 pairs per round



(d) NB Jaccard 10 pairs per round (e) NB Jaccard 20 pairs per round (f) NB Jaccard 50 pairs per round



(g) NB N-W 10 pairs per round (h) NB N-W 20 pairs per round (i) NB N-W 50 pairs per round

Selection Strategies

- Uncertainty performs worse than random or representative in our setting
- Representative performs similar to Random

- The representative strategy, taking into account items from the entire pool of unlabeled instances performs best

Conclusions

- Proposed a flexible framework for active learning from the crowd
- Tested on the particular scenario of duplicates detection

- When employing such a framework the choice of automatic method is very important as it guides the acquisition of new labels
- An optimal resource allocation schema has to be found, as after a certain point, spending extra will not provide better performance
- Such frameworks are sensible to the quality of crowd data, and analyzing the worker behavior is a prerequisite
- The Selection Strategy plays a crucial role; a representative strategy gives better results than one based on uncertainty

Future directions

- Direct extension: use the crowd to learn how to create a merged representation of the detected duplicates

- Experiment with other types of tasks and data
- Employ various crowd label aggregation strategies and worker reliability estimation
- Investigate the influence of agreement on performance
- In depth study on Selection Strategies

Thank you!

Q&A