HOCHSCHULE
MITTWEIDA
University of Applied Sciences

IMMM 2020

Towards Inter-Rater-Agreement-Learning

Michael Spranger

How AI shape our Life

It is difficult to think of a major industry that AI will not transform. This includes healthcare, education, transportation, retail, communications, and agriculture. There are surprisingly clear paths for AI to make a big difference in all of these industries.
Andrew Ng

It's very clear that AI is going to impact every industry. I think that every nation needs to make sure that AI is a part of their national strategy. Every country will be impacted.
Jensen Huang

I think that AI will lead to a low cost and better quality life for millions of people. Like electricity, it's a possibility to build a wonderful society.
Andrew Ng

A good AI's needs

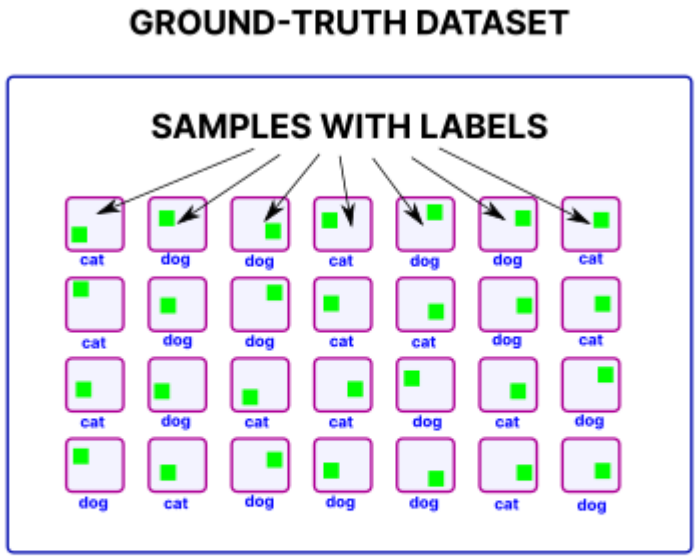
Commonly, a human-labeled dataset is considered as ground-truth

The truth is rarely pure and never simple.
Oscar Wilde

Ideally, an expert-labeled dataset should be considered as ground-truth

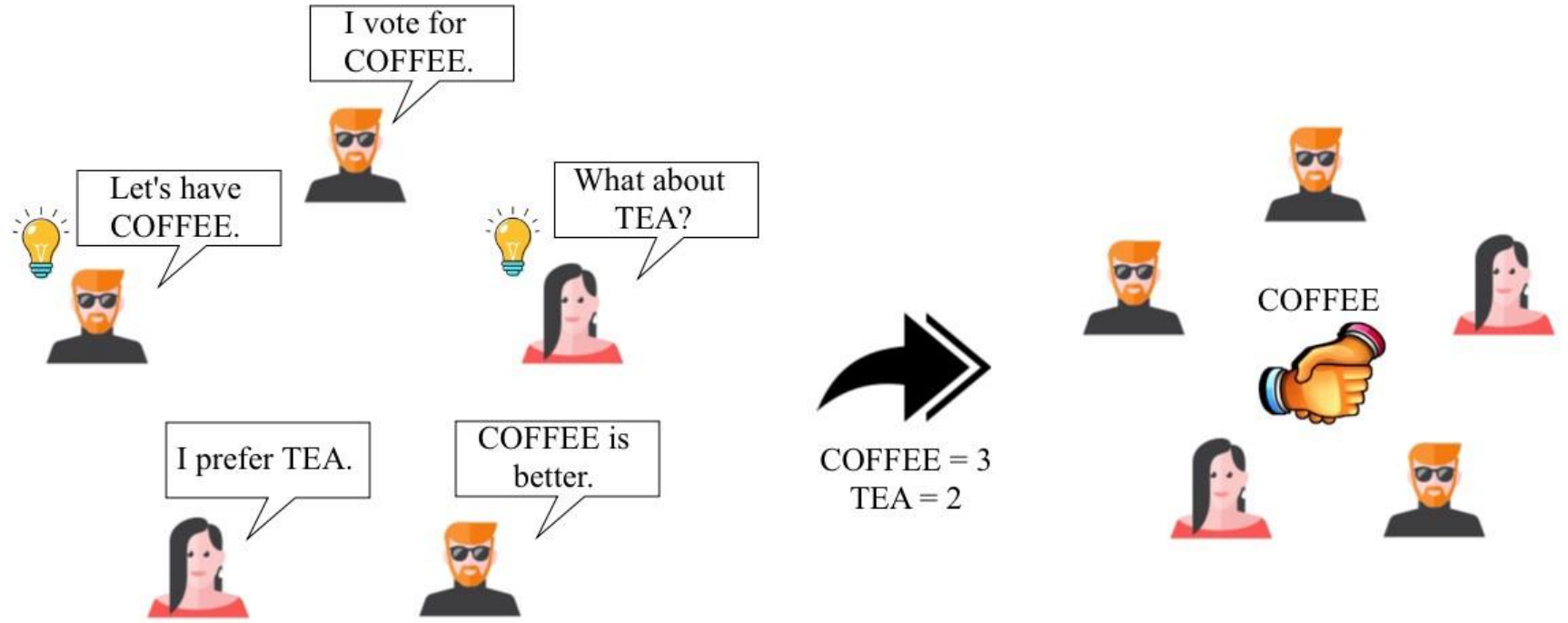
If you have a lot of data and you want to create value from that data, one of the things you might consider is building up an AI team.
Andrew Ng

Usually, machine learning needs much data, but there are not enough experts to label it



https://wao.ai/blog/dataset-vs-ground-truth-dataset

Consensus by majority



https://www.google.com/search?q=consensus&tbm=isch&hl=de&hl=de&tbs&client=opera&hs=LFj&sa=X&ved=0CAEQpwVqFwoTCMisp7uSxesCFQAAAAAdAAAAABAR&biw=1849&bih=929#imgrc=L8Lr0YwMhqTPuM

Are all ratings equally valuable?

Rater‘s dimensions

?

==

Rater‘s dimensions

already labeled items

1 — 1000 — 2000

Should all ratings for an item have the same weight?

Weighted Learning Approach

$\boldsymbol{R_j}$	rater's competence
$\boldsymbol{f(t_{ji}, C_j)}$	value depending on the response time and the conscientiousness of a specific rater who needs to annotate an item i at time t
$\boldsymbol{\beta_i}$	weighting parameter
$\boldsymbol{x_i}$	feature (self-judgement, Intra-Rater-Agreement,…)

$$w_{ji} = R_j - f(t_{ji}, C_j)$$

$$R_j = \frac{\sum_{l=1}^{n} \beta_l x_{lj}}{\frac{1}{|J|} \sum_{j'=1}^{|J|} \sum_{l=1}^{n} \beta_l x_{lj'}} \quad - \quad f(t_{ji}) = \begin{cases} 0, & C_j > \bar{t}_i \wedge t_{ji} \in [\bar{t}_i, C_j] \\ 0, & C_j < \bar{t}_i \wedge t_{ji} \in [C_j, \bar{t}_i] \\ t_{ji} - \bar{t}_i, & C_j > \bar{t}_i \wedge t_{ji} \notin [\bar{t}_i, C_j] \\ \bar{t}_i - t_{ji}, & C_j < \bar{t}_i \wedge t_{ji} \notin [C_j, \bar{t}_i] \end{cases}$$

Weighted Learning Approach

Taking time (already labeled items) into account:

$$w_{ji} = (1 - b_{ji})[R_j - f(t_{ji}, C_j)] + \frac{1}{b_{ij}} \sum w_{j(i-1)}$$

$$b_{ji} = \frac{1}{\lambda(i-1)} \sum_{k=0}^{i-1} \begin{cases} 1, & for\ j\ in\ majority\ for\ item\ k \\ 0, & else \end{cases}$$

Preliminary Results

3000 texts from music domain

Threshold	is Music		Uncertain		not Music		No Majority	
	UW	W	UW	W	UW	W	UW	W
0.5	1534	1560	3	6	1344	1357	119	77
0.55	1456	1481	2	2	1298	1302	244	215
0.6	1373	1410	2	1	1240	1249	386	339
0.65	1278	1314	0	0	1175	1182	547	504
0.7	1128	1191	0	0	1115	1139	757	670
0.75	982	1072	0	0	1064	1081	954	847
0.8	825	918	0	0	998	1025	1177	1057
0.85	648	739	0	0	931	952	1421	1309
0.9 0	435	504	0	0	821	842	1744	1654
0.95	236	270	0	0	601	636	2163	2094

„No Majority“ decreases for each threshold

Conclusion & Future Work

- flexible weighting approach for Inter-Rater-Agreement
- strengths and weaknesses of different raters are considered
- automatic adaptation to dynamic user characteristics like concentration, motivation etc.
- results on a first dataset providing only few parameters leads to less items with “no majority”
- Future work will incorporating tests on a multi-lingual dataset including more features