ULF: Unsupervised Labeling Function Correction using Cross-Validation for Weak Supervision

Anastasiia Sedova, Benjamin Roth
{anastasiia.sedova, benjamin.roth}@univie.ac.at
University of Vienna

Paper on ArXiv:

Motivation

Weak supervision: the model is trained on the data, which was automatically labeled using heuristics (key words, external KB, etc) formulated as labeling functions. Some labeling functions (LFs) correctly captures some samples but mislabels others.

- Example: Lf **"my"** assigns SPAM class.
  - *Subscribe to **my** channel, check **my** channel out* -> SPAM ✔
  - *It looks so real and **my** daughter is a big fan and …* -> SPAM ✘
- Such hard LFs to class assignments often results in a tie and incorrect assignment.

	Sample	Matched Labeling Functions	Assigned Label
1	**CHECK MY** CHANNEL **OUT PLEASE**. I DO SINGING COVERS	keyword_my, keyword_please regex_check_out	SPAM
2	Hello! I'm Marian … I wanted to play my own pop and pop-rock **songs**. It would mean a lot if you could have a look at **my** channel … if u like, **subscribe** to it! XOXO THANKS!!	keyword_my, keyword_subscribe keyword_song textblob_subjectivity	SPAM/ HAM
3	It looks so real and **my** daughter is a big fan and she likes a lot of your **songs**.	keyword_my short_songs	SPAM/ HAM
4	Follow me on Twitter @mscalifornia95	no matches	—

We propose to use a fine-adjusted LFs to class assignment in order to correct the label mistakes.

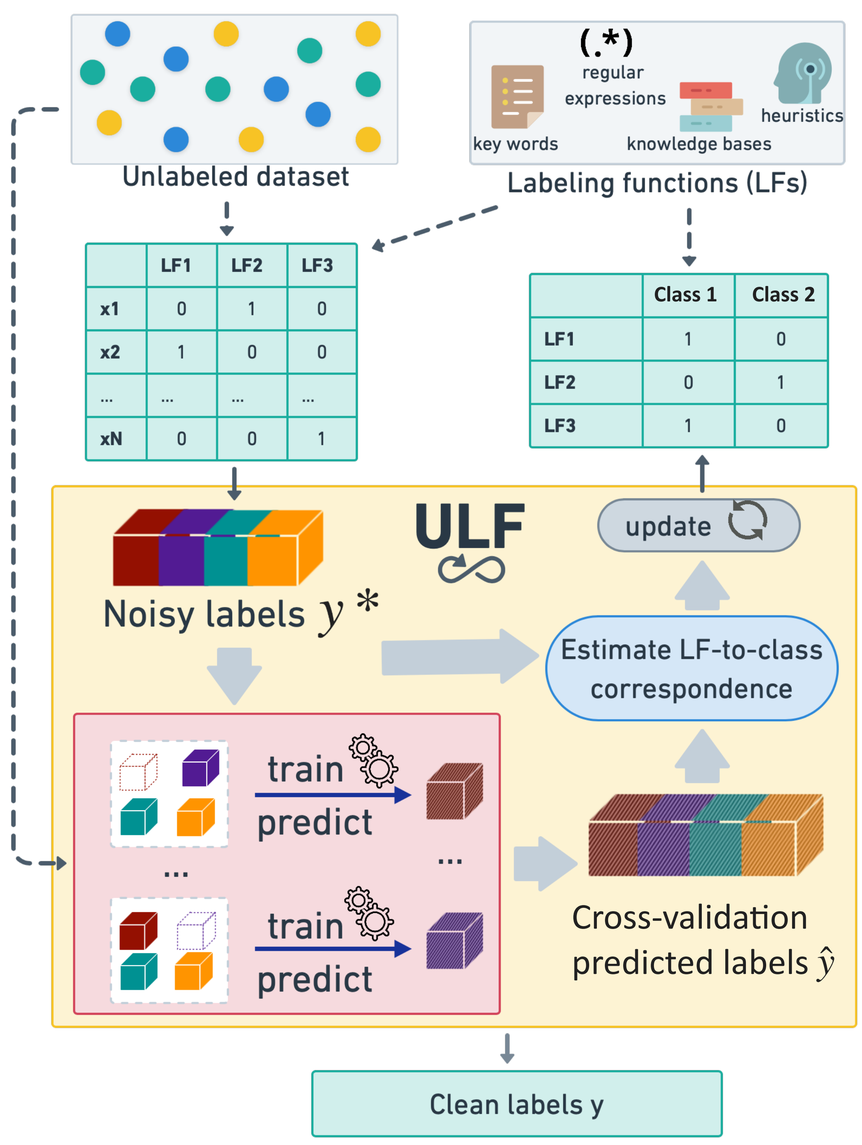
ULF tunes the joint distribution between **LFs** and **labels** based on highly confident class cross-validation predictions and their cooccurrence with LFs.

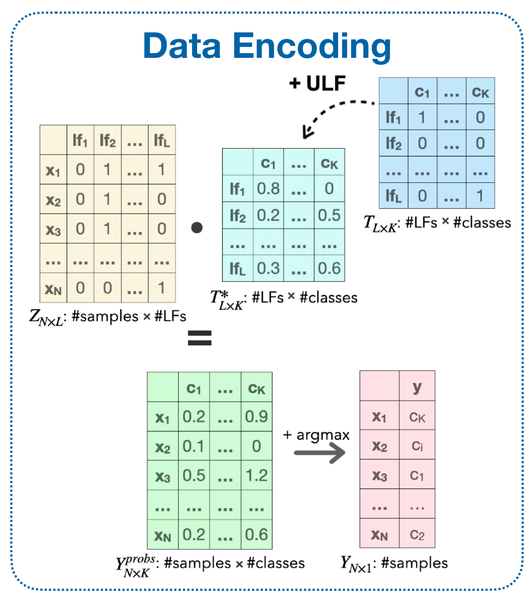
Cross-Validation Component

- Split the data into *k* folds according to matched LFs (i.e., *signatures*).
- Train *k* models on *k-1* folds.
- Apply the models to the held-out folds to calculate *cross-validation predicted probabilities.*
- Convert the probabilities into **labels** $\hat{y}$ w.r.t. class average thresholds t_j:

$$t_j := \frac{\sum_{x_i \in X_{\tilde{y}=j}} p(\tilde{y}=j; x_i, \theta)}{|X_{\tilde{y}=j}|}$$

Intuition: A mismatch between predictions of a model trained on a large portion of the LFs and labels generated by held-out LFs can indicate *noise specific to the held-out LFs*.





- A confidence matrix $C_{L\times K}$ estimates the joint distribution between matched LFs and predicted labels:

$$C_{l_i,\hat{y}_j} = |\{x_i \in X : \hat{y}_i = \tilde{y}_j, l_i \in L_{x_i}\}|$$

- Its calibrated version $Q_{L\times K}$ (sums up to the total number of training samples, and the sum of counts for each LF is the same as in the original Z matrix) is used to re-estimate T:

$$T* = p * \hat{Q} + (1-p) * T$$

Experiments

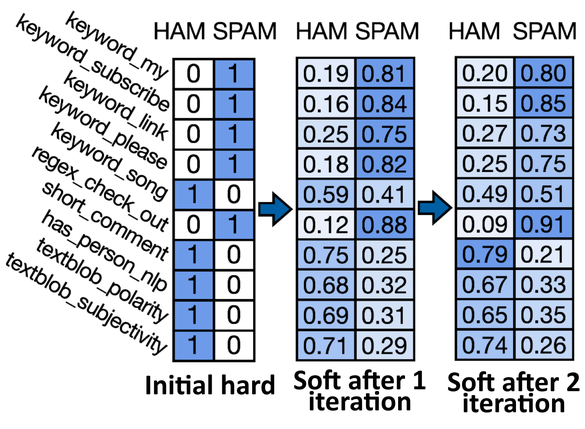
Experimental Results

	YouTube	Spouse	TREC	SMS	Yorùbá	Hausa	Avg
Gold	98.8	-	96.6	97.7	67.3	83.5	88.8
Majority Vote (MV)	93.2	21.3	68.6	93.0	48.1	43.9	61.4
MeTaL (Ratner et al., 2019)	96.0	19.6	55.8	89.2	58.6	41.6	60.1
Snorkel-DP (Ratner et al., 2020)	95.6	32.6	61.8	94.6	58.7	45.7	64.8
FlyingSquid (Fu et al., 2020)	94.0	14.9	35.8	23.7	32.4	45.1	41.0
WeaSEL (Cachay et al., 2021)	96.0	14.9	64.4	23.6	49.6	43.2	48.6
FABLE (Zhang et al., 2023)	94.8	27.8	54.6	91.1	23.2	18.6	51.7
MV + Cosine (Yu et al., 2021)	96.4	33.3	65.8	93.6	52.6	45.4	64.5
MeTaL + Cosine	95.6	26.9	67.4	80.7	**62.0**	45.5	63.0
Snorkel-DP + Cosine	96.0	28.1	73.8	96.1	55.0	46.5	65.9
FlyingSquid + Cosine	95.6	24.9	38.6	90.1	33.3	41.5	54.0
FABLE + Cosine	94.0	33.9	70.6	**97.7**	60.1	44.7	66.8
ULF (Ours)	**96.8**	**36.9**	**76.8**	96.2	55.8	**48.2**	**68.4**

- **WS Datasets:** 4 English + 2 African languages
- **Tasks:** sentiment analysis, relation extraction, topic classification
- **Models**: RoBERTa/multilingual BERT (+ optional Cosine training step)
- **Unlabeled samples:** λ samples are randomly labeled and included in cross-validation training + reestimated in next iterations.

We also conducted the experiments with feature-based models - check out the paper!

T matrix transformation with ULF



- **Parameters** (tuned on the manually labeled validation data): the usual model training parameters + #ULF iterations, #ULF folds, multiplying coefficient *p*, soft/hard labels, non-labeled data rate λ.

CV / Final	YouTube	Spouse	TREC	SMS	Yorùbá	Hausa
FT_FT	96.8	22.0	68.2	96.1	54.6	43.0
FT_Cos	94.4	36.9	76.6	96.2	54.2	48.2
Cos_FT	95.2	21.3	68.6	96.1	55.8	43.6
Cos_Cos	94.8	33.0	76.8	96.1	54.2	44.5

Implemented within the knodle framework

	Sample	LFs matched	Noisy Label	Corrected Label	Gold Label
1	Hello! I'm Marian … I wanted to play my own pop and pop-rock **songs**. It would mean a lot if you could have a look at **my** channel … if u like, **subscribe** to it! XOXO THANKS!!	keyword_my keyword_subscribe keyword_song textblob_subjectivity	HAM	SPAM	SPAM
2	Nice **song** .See **my** new track.	keyword_my keyword_song textblob_subjectivity	HAM	SPAM	SPAM
3	'HAPPY BIRTHDAY KATY :) **http://giphy.com/gifs/birthday-flowers-happy-gw3JY2uqiaXKaQXS/fullscreen** …	keyword_link textblob_subjectivity	HAM	SPAM	HAM
4	The little PSY is suffering Brain Tumor and only has 6 more months to live. **Please** pray to him and the best lucks.	keyword_please textblob_subjectivity	HAM	SPAM	HAM
5	Follow me on Twitter @mscalifornia95	—	HAM	SPAM	SPAM