Data Programming using Continuous and Quality-Guided Labeling Functions

Oishik Chatterjee, **Ganesh Ramakrishnan**, Sunita Sarawagi

July 15, 2020

Department of Computer Science and Engineering, IITB

Labeling Functions on ‘Spouse’ Extraction Task [Bach *et. al.*, 2017]

SpouseDict = {’spouse’, ’married’, wife’, ’husband’, ’ex-wife’, ’ex-husband’}

FamilyDict = {’father’, ’mother’, ’sister’, ’brother’, ’son’, ’daughter’,’grandfather’, ’grandmother’, ’uncle’, ’aunt’,’cousin’ }$\bigotimes${+”,+’-in-law’}

OtherDict = {’boyfriend’, ’girlfriend’ ’boss’, ’employee’, ’secretary’, ’co-worker’}

SeedSet = {(’Barack Obama’, ’Michelle Obama ’), (’Jon Bon Jovi’,’Dorothea Hurley’),(’Ron Howard’,’Cheryl Howard’),.....}

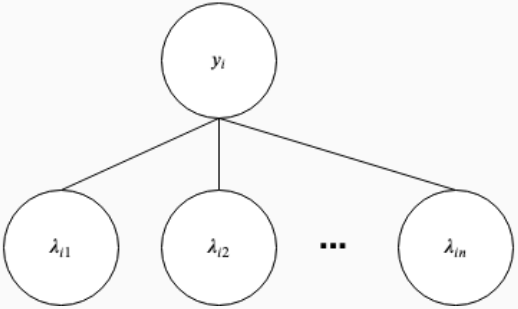
Id	Description
LF1	If some word in **SpouseDict** is present between E_1 and E_2 or within 2 words of either, return 1 else return 0
LF2	If some word in **FamilyDict** is present between E_1 and E_2, return -1 else return 0.
LF3	If some word in **OtherDict** is present between E_1 and E_2, return -1 else return 0.
LF4	If both E_1 and E_2 occur in **SeedSet**, return 1 else return 0.
LF5	If the number of word tokens lying between E_1 and E_2 are less than 4, return 1 else return 0.

Table 1: Discrete labeling functions (LFs) based on dictionary lookups or thresholded distance for the *spouse* relationship extraction task

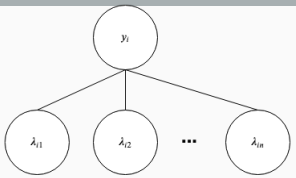
Problem Setting

Motivation: Lack of labeled data. Human designed labeling functions (LFs) assigning noisy labels to instances. Use these labels to generate labeled data. *(Data Programming)*
Problem Statement: Learn a generative model of true label distributions over LFs.
Data Programmming using Snorkel (Bach et al. 2017):

- Discrete Probabilistic Graphical Model
- Shared Parameters for agreement and disagreement of labeling functions



Problem Setting: Data Programmming using Snorkel *[Bach et al. 2017)*



Limitations:

1. Training instability due to unsupervised nature of the problem.
2. Highly sensitive to initialization, number of epochs, learning rate, *etc.*
3. Does not support continuous labeling functions.

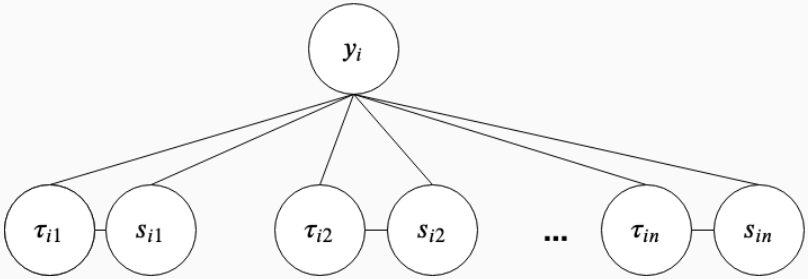
Joint probability distribution of y (true label for an instance(**X**) and $\Lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{in})$ (labels assigned by the n labeling functions) is

$$P_\theta(y, \Lambda_i) = \frac{1}{Z_\theta} \exp\left(\sum_{j=1}^{n} \phi_j(y, \lambda_{ij})\right)$$

$$\phi_j(y, \lambda_{ij}) = \begin{cases} \theta_{jy} & \text{if } \lambda_{ij} = k_j, \\ -\theta_{jy} & \text{if } \lambda_{ij} \neq k_j, \lambda_{ij} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Problem Setting: Our Solution - CAGE *[AAAI 2020]*

CAGE stands for **C**ontinuous **A**nd quality **G**uided lab**E**ling functions.



Given n labeling functions $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ which can be **continuous** or discrete.

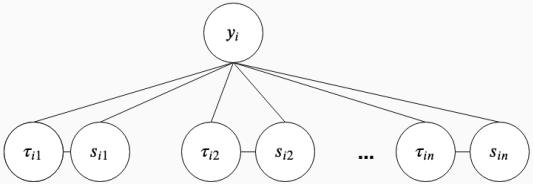
- Each LF λ_j, for a given instance x_i,
  - outputs a label $\tau_{ij} = k_j$ if triggered
  - else outputs $\tau_{ij} = 0$
- Further, if λ_j is a continuous LF, it also outputs a score $s_{ij} \in (0, 1)$.

For each x_i, we model the joint probability of the true label and the (labels, scores)

Problem Setting: Continuous LFs in CAGE

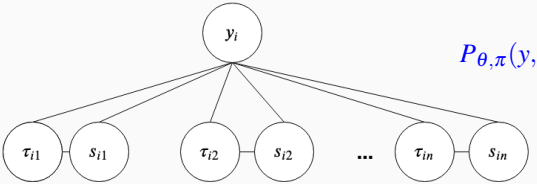
Id	Class	Description
LF1	+1	$\max$ [cosine(word-vector(u), word-vector(v))-0.8]$_+$: $u \in$ **SpouseDict** and $v \in$ {words between E_1, E_2}.
LF2	-1	$\max$ [cosine(word-vector(u), word-vector(v))-0.8]$_+$: $u \in$ **FamilyDict** and $v \in$ {words between E_1, E_2}.
LF3	-1	$\max$ [cosine(word-vector(u), word-vector(v))-0.8]$_+$: $u \in$ **OtherDict** and $v \in$ {words between E_1, E_2}.
LF4	-1	$\max$ [0.2 - Norm-Edit-Dist(E_1, E_2, u, v)]$_+$: $(u,v),(v,u) \in$ **SeedSet**.
LF5	+1	[1 - (number of word tokens between E_1 and E_2)/5.0]$_+$

CAGE: The Probabilistic Model



Joint probability distribution of y (true label for an instance(**X**) and $\Lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{in})$ (labels assigned by the n labeling functions) and their scores $(s_{i1}, s_{i2}, \ldots, s_{in})$ is .....

CAGE: The Probabilistic Model



$$P_{\theta,\pi}(y,\tau_i,s_i) = \frac{1}{Z_\theta}\prod_{j=1}^{n}\psi_\theta(\tau_{ij},y)\left(\psi_\pi(\tau_{ij},s_{ij},y)\right)^{\text{cont}(\lambda_j)}$$

Joint probability distribution of y (true label for an instance(**X**) and $\Lambda_i = (\lambda_{i1},\lambda_{i2},\ldots,\lambda_{in})$ (labels assigned by the n labeling functions) and their scores $(s_{i1},s_{i2},\ldots,s_{in})$ is .....

$$\psi_\theta(\tau_{ij},y) = \begin{cases} \exp(\theta_{jy}) & \text{if } \tau_{ij} \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

$$\psi_\pi(\tau_{ij},s_{ij},y) = \begin{cases} \textit{Beta}(s_{ij};\alpha_a,\beta_a) & \text{if } k_j = y \;\&\; \tau_{ij} \neq 0, \\ \textit{Beta}(s_{ij};\alpha_d,\beta_d) & \text{if } k_j \neq y \;\&\; \tau_{ij} \neq 0, \\ 1 & \text{otherwise} \end{cases}$$

Note $\alpha_a = q_j^c\pi_{jy}$ and $\beta_a = (1-q_j^c)\pi_{jy}$ are parameters of the agreement distribution and $\alpha_d = (1-q_j^c)\pi_{jy}$ and $\beta_d = q_j^c\pi_{jy}$ are parameters of the disagreement distribution, where π_{jy} is constrained to be strictly positive. To impose $\pi_{jy} > 0$ while also maintaining differentiability, we reparametrize π_{jy} as $\exp(\rho_{jy})$.

$$Z_\theta = \sum_y \prod_j \sum_{\tau_j \in \{k_j,0\}} \psi_\theta(\tau_j,y)\int_{s_j=0}^{1}\psi_\pi(\tau_j,s_j,y) = \sum_{y\in\mathscr{Y}}\prod_j(1+\exp(\theta_{jy}))$$

Relationship with Snorkel

CAGE Model Potential

$$\psi_\theta(\tau_{ij}, y) = \begin{cases} \exp(\theta_{jy}) & \text{if } \tau_{ij} \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

Setting $\theta_{j,+1} = -\theta_{j,-1}$ in the CAGE model .....

Relationship with Snorkel

CAGE Model Potential

$$\psi_\theta(\tau_{ij}, y) = \begin{cases} \exp(\theta_{jy}) & \text{if } \tau_{ij} \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

Snorkel Model Potential

$$\phi_j(y, \lambda_{ij}) = \begin{cases} \theta_{jy} & \text{if } \lambda_{ij} = j, \\ -\theta_{jy} & \text{if } \lambda_{ij} \neq j, \lambda_{ij} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Setting $\theta_{j,+1} = -\theta_{j,-1}$ in the CAGE model .....

Few simplifications in CAGE lead to the snorkel model

- Coupling of θ_{jy} parameters in $\phi_j(y, \lambda_{ij})$
- Not including continuous LFs and the associated potentials $\left(\psi_\pi(\tau_{ij}, s_{ij}, y)\right)^{\text{cont}(\lambda_j)}$
- Ignoring quality guides q_j^t (next)...

CAGE: The Training Objective

$$\max_{\theta,\pi} LL(\theta,\pi|D) + R(\theta,\pi|\{q_j^t\})$$

CAGE: The Training Objective

$$\max_{\theta,\pi} LL(\theta,\pi|D) + R(\theta,\pi|\{q_j^t\})$$

WHY $R(\theta,\pi|\{q_j^t\})$? Unsupervised likelihood training inherently unstable

$$LL(\theta,\pi|D) = \sum_{i=1}^{m} \log \sum_{y\in\mathcal{Y}} P_{\theta,\pi}(\tau_i, s_i, y)$$

$$= \sum_{i=1}^{m} \log \sum_{y\in\mathcal{Y}} \prod_{j=1}^{n} \psi_j(\tau_{ij}, y) \left(\psi_j(s_{ij}, \tau_{ij}, y)\right)^{\text{cont}(\lambda_j)} - m \log Z_\theta$$

Recall, from the previous slide,

$$P_{\theta,\pi}(y, \tau_i, s_i) = \frac{1}{Z_\theta} \prod_{j=1}^{n} \psi_\theta(\tau_{ij}, y) \left(\psi_\pi(\tau_{ij}, s_{ij}, y)\right)^{\text{cont}(\lambda_j)}$$

CAGE: The Training Objective

$$\max_{\theta,\pi} LL(\theta,\pi|D) + R(\theta,\pi|\{q_j^t\})$$

WHY $R(\theta,\pi|\{q_j^t\})$? Unsupervised likelihood training inherently unstable

WHAT is $R(\theta,\pi|\{q_j^t\})$? Options:

- Match learned joint distribution $P_{\theta,\pi}$ of y and τ_j with user-provided quality guides q_j^t
  - For continuous LFs parameterize Beta distribution to combine quality guides and learning.
  - Empowers programmer to stabilize training via easy quality guides q_j^t on a LF (e.g. accuracy ≥ 0.5)

$$LL(\theta,\pi|D) = \sum_{i=1}^{m} \log \sum_{y\in\mathscr{Y}} P_{\theta,\pi}(\tau_i, s_i, y)$$

$$= \sum_{i=1}^{m} \log \sum_{y\in\mathscr{Y}} \prod_{j=1}^{n} \psi_j(\tau_{ij}, y) \left(\psi_j(s_{ij}, \tau_{ij}, y)\right)^{\text{cont}(\lambda_j)} - m \log Z_\theta$$

Recall, from the previous slide,

$$P_{\theta,\pi}(y, \tau_i, s_i) = \frac{1}{Z_\theta} \prod_{j=1}^{n} \psi_\theta(\tau_{ij}, y) \left(\psi_\pi(\tau_{ij}, s_{ij}, y)\right)^{\text{cont}(\lambda_j)}$$

CAGE: The Training Objective

$$\max_{\theta,\pi} LL(\theta,\pi|D) + R(\theta,\pi|\{q_j^t\})$$

$$LL(\theta,\pi|D) = \sum_{i=1}^{m} \log \sum_{y\in\mathscr{Y}} P_{\theta,\pi}(\tau_i, s_i, y)$$

$$= \sum_{i=1}^{m} \log \sum_{y\in\mathscr{Y}} \prod_{j=1}^{n} \psi_j(\tau_{ij}, y) \left(\psi_j(s_{ij}, \tau_{ij}, y)\right)^{\text{cont}(\lambda_j)} - m \log Z_\theta$$

Recall, from the previous slide,

$$P_{\theta,\pi}(y, \tau_i, s_i) = \frac{1}{Z_\theta} \prod_{j=1}^{n} \psi_\theta(\tau_{ij}, y) \left(\psi_\pi(\tau_{ij}, s_{ij}, y)\right)^{\text{cont}(\lambda_j)}$$

WHY $R(\theta,\pi|\{q_j^t\})$? Unsupervised likelihood training inherently unstable

WHAT is $R(\theta,\pi|\{q_j^t\})$? Options:

- Match learned joint distribution $P_{\theta,\pi}$ of y and τ_j with user-provided quality guides q_j^t
  - For continuous LFs parameterize Beta distribution to combine quality guides and learning.
  - Empowers programmer to stabilize training via easy quality guides q_j^t on a LF (e.g. accuracy ≥ 0.5)
- Other options considered:
  - Sign penalty on raw parameters to favor agreement
  - Constraints on LF accuracy calculated on data

CAGE: The Training Objective & Constraints $R(\theta, \pi | \{q_j^t\})$

$$\max_{\theta,\pi} LL(\theta, \pi | D) + R(\theta, \pi | \{q_j^t\})$$

Matching learned joint distribution $P_{\theta,\pi}$ of y and τ_j with user-provided quality guides q_j^t :

$$LL(\theta, \pi | D) = \sum_{i=1}^{m} \log \sum_{y \in \mathscr{Y}} P_{\theta,\pi}(\tau_i, s_i, y)$$

$$= \sum_{i=1}^{m} \log \sum_{y \in \mathscr{Y}} \prod_{j=1}^{n} \psi_j(\tau_{ij}, y) \left(\psi_j(s_{ij}, \tau_{ij}, y)\right)^{\text{cont}(\lambda_j)} - m \log Z_\theta$$

$$R(\theta | \{q_j^t\}) = \sum_j q_j^t \log P_\theta(y = k_j | \tau_j = k_j)$$

$$+ (1 - q_j^t) \log(1 - P_\theta(y = k_j | \tau_j = k_j))$$

Recall, from the previous slide,

$$P_{\theta,\pi}(y, \tau_i, s_i) = \frac{1}{Z_\theta} \prod_{j=1}^{n} \psi_\theta(\tau_{ij}, y) \left(\psi_\pi(\tau_{ij}, s_{ij}, y)\right)^{\text{cont}(\lambda_j)}$$

$$P_\theta(y = k_j | \tau_j = k_j) = \frac{P_\theta(y = k_j, \tau_j = k_j)}{P_\theta(\tau_j = k_j)}$$

$$= \frac{\mathsf{M}_j(k_j) \prod_{r \neq j} (1 + \mathsf{M}_r(k_j))}{\sum_{y \in \mathscr{Y}} \mathsf{M}_j(y) \prod_{r \neq j} (1 + \mathsf{M}_r(y))}$$

Data independent, Efficient, and Accurate!

Experimental Setup

Datasets:

1. Spouse: Relation extraction dataset - label candidate pairs of entities in a sentence as expressing a ‘spouse’ relation or not
2. Spam SMS: Binary spam/no-spam classification dataset with 5574 documents: 3700 unlabeled-train and 1872 labeled-test
3. CDR: Relation extraction dataset where the task is to detect whether or not a sentence expresses a ‘chemical cures disease’ relation
4. Dedup: 32 thousand pairs of noisy citation records with fields like Title, Author, Year etc. Task - detect if record pairs are duplicates
5. Ionosphere
6. Iris

Training Setup:

1. $q_{tj}(Discrete) = 0.9$
2. $q_{cj}(Continuous) = 0.85$
3. Learning rate = 0.001
4. Epochs = 100

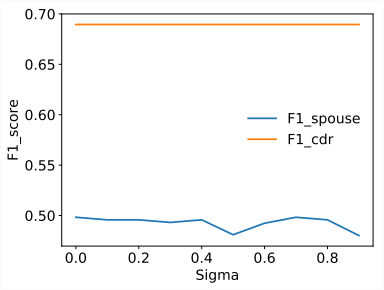
Overall Results

	Datasets					
	Spouse	CDR	SMS	Ion	Iris	Dedup
Majority	0.17	0.53	0.23	0.79	0.84	-
Snorkel	0.41	0.66	0.34	0.70	0.87	-
CAGE_{-C-G}	0.48	0.69	0.34	0.81	0.87	-
CAGE_{-C}	0.50	0.69	0.45	0.82	0.87	-
CAGE	**0.58**	**0.69**	**0.54**	**0.97**	**0.87**	**0.79**

Overall Results (F1) with predictions from various generative models contrasted with the Majority baseline.

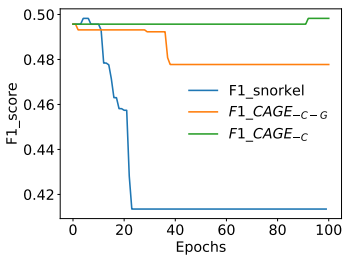
Quality Guides

	Spouse	CDR	Sms	Ion
CAGE$_{-C-G+-P}$	0.48	0.69	0.34	0.81
CAGE$_{-C-G}$	0.48	0.69	0.34	0.81
CAGE$_{-C,dataG}$	0.48	0.69	0.34	0.81
CAGE$_{-C}$	**0.50**	**0.69**	**0.45**	**0.82**

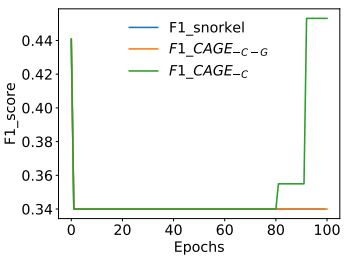


F1 with increasing distortion in the guess of the LF quality guide, q_j^t.

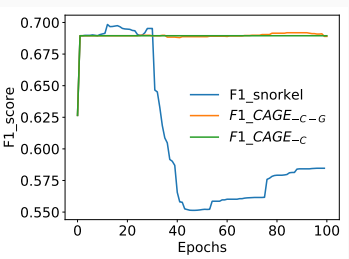
Quality Guides & Performance wrt epochs



(a) Spouse



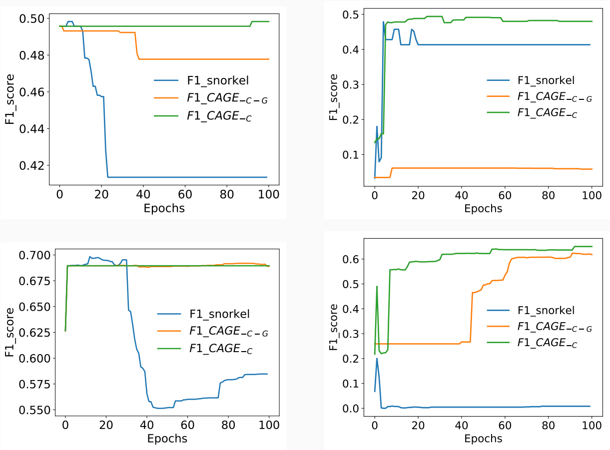
(b) SMS



(c) CDR

F1 with increasing number of training epochs compared across snorkel, CAGE_{-C-G} and CAGE_{-C}, for three datasets. For each dataset, in the absence of guides, we observe unpredictable variation in test F1 as training progresses.

Sensitivity to initialization



(a) Agreeing initialization (b) Random Initialization.

F1 with increasing number of training epochs compared across snorkel, CAGE$_{-C-G}$ and CAGE$_{-C}$, for two datasets: Spouse (top-row) and CDR(bottom-row). CAGE$_{-C}$ is able to recover from any initialization whereas methods without guides fare even worse with random initialization.

Problem Setting: Summarily

- **Continuous** LFs improve recall.
- Snorkel's unsupervised likelihood training is inherently unstable.
- CAGE allows **quality guides** to stabilize learning.
- Elegant method of incorporating guides into likelihood training.