REAL: A Representative Error-Driven Approach for Active Learning

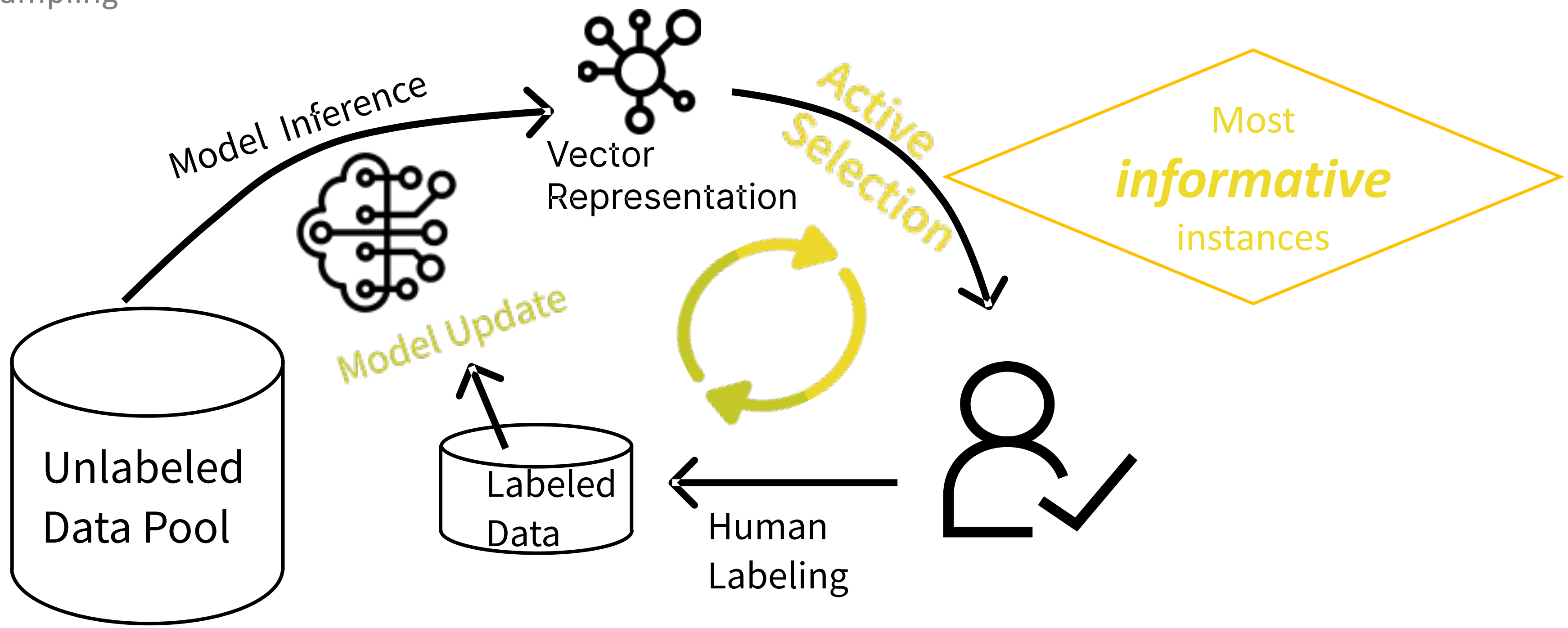
Cheng Chen[1,2], Yong Wang[2], Lizi Liao[2], Yueguo Chen[1], Xiaoyong Du[1]

中国人民大學
RENMIN UNIVERSITY OF CHINA

SMU
SINGAPORE MANAGEMENT UNIVERSITY

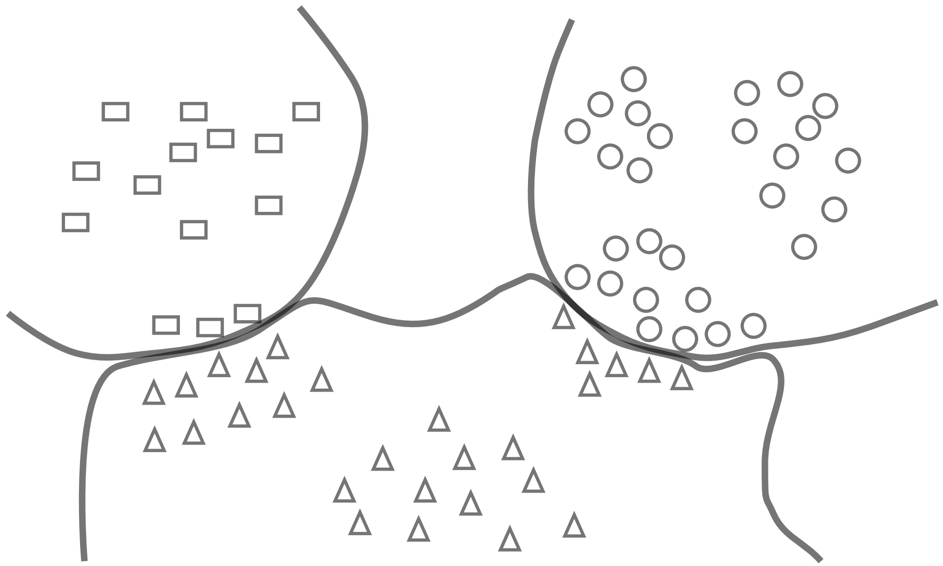
Active Learning (AL)

Pool-based sampling



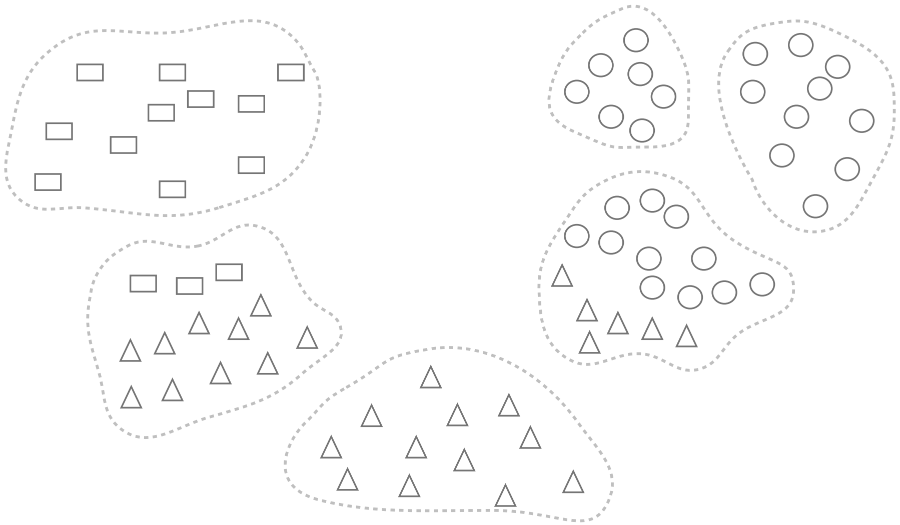
Background-Uncertainty

Uncertainty-based AL selects the most uncertain instances for the model.



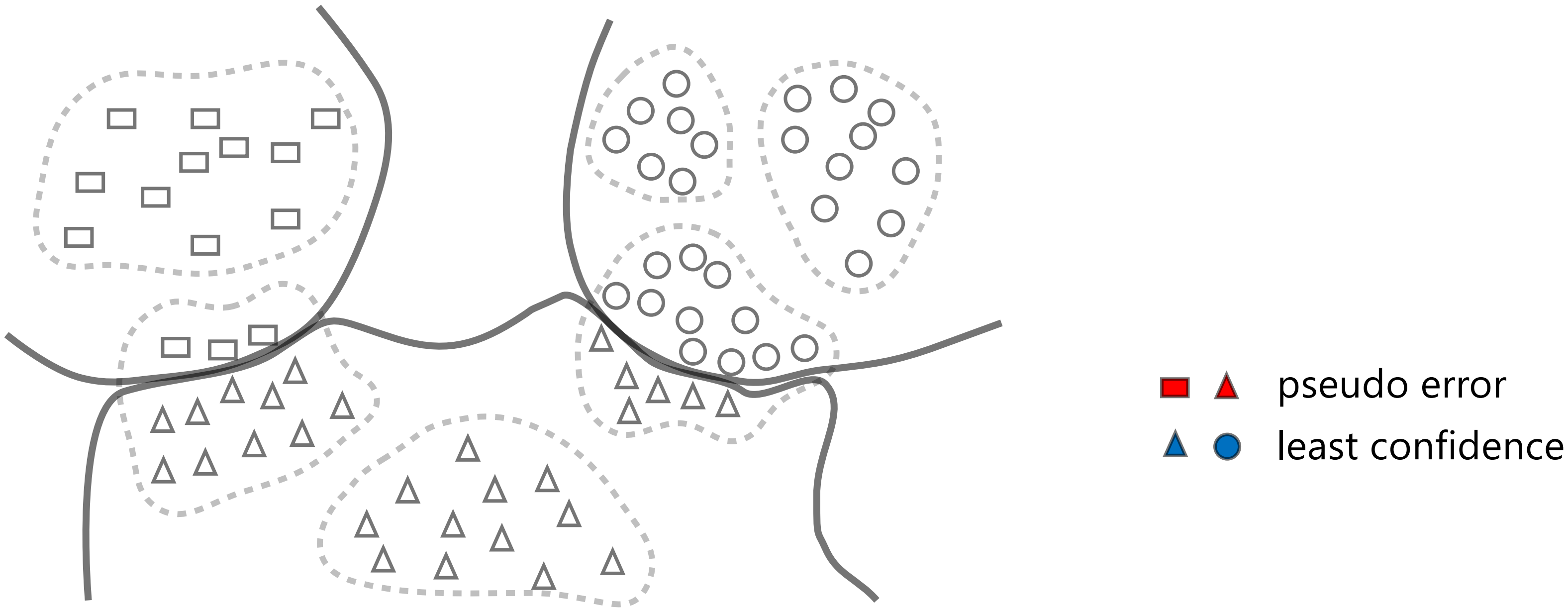
Background-Diversity

Diversity-based AL aims to maximize the diversity of sampled instances.



Motivation-REAL

Erroneous instances are more informative for AL [1,2].
REAL selects *representative errors* near decision boundary.

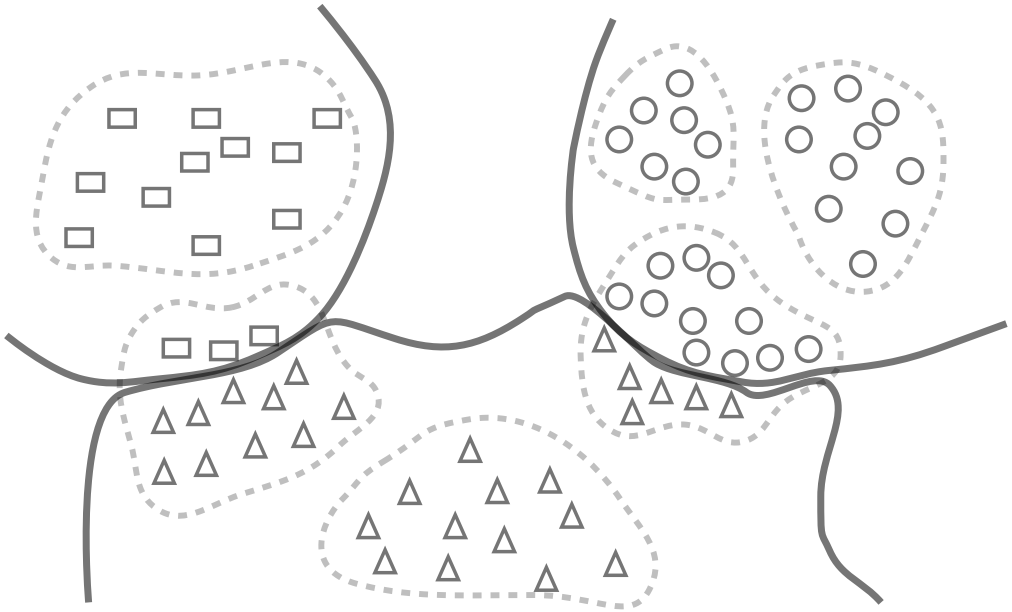


[1] Choi et al., Vab-al: Incorporating class imbalance and difficulty with variational bayes for active learning, CVPR'2021
[2] Krempl et al., Optimised probabilistic active learning (opal) for fast, non-myopic, cost-sensitive active classification. ML'2015

Contributions

- REAL: a new AL sampling algorithm dedicated to representative errors.
- New SOTA result on five text classification benchmarks.
- Insights on error distribution:
  - most errors are along the decision boundary;
  - REAL's active selections align well with that of ground-truth errors.

REAL: Representative Error-Driven Active Learning



- K-Means clustering
- Assign pseudo labels
- Find pseudo errors
- Add least confidence

REAL - Pseudo Error Identification

- The predicted label for an individual instance:

$$\widetilde{y}_i = \underset{j \in \{1,\ldots,Y\}}{\operatorname{argmax}} \left[\mathcal{M}(\mathrm{x}_i; \theta^{(t)})\right]_j$$

- The pseudo label of cluster:

$$y_{maj} = \underset{j}{\operatorname{argmax}} \Big(\sum_{i \in \mathcal{C}_k^{(t)}} \mathbb{1}\{\widetilde{y}_i = j\} \Big) / |\mathcal{C}_k^{(t)}|$$

- The instances that are not predicted as y_{maj} are defined as pseudo errors in the corresponding cluster $\mathcal{C}_k^{(t)}$.

REAL - Adaptive Sampling

- Goal: adaptive sampling of representative errors
- Single instance's erroneous probability:

$$\epsilon(\mathrm{x}_e) = 1 - [\mathcal{M}(\mathrm{x}_e; \theta^{(t)})]_{maj}$$

- The density of pseudo errors ϵ_k for cluster $\mathcal{C}_k^{(t)}$:

$$\epsilon_k = \sum \epsilon(\mathrm{x}_e)$$

- The sampling budget b_k for the cluster $\mathcal{C}_k^{(t)}$:

$$b_k = \left\lfloor b \frac{\epsilon_k}{\sum_i \epsilon_i} \right\rfloor, \forall k \in \{1 \dots K\}$$

Experiments

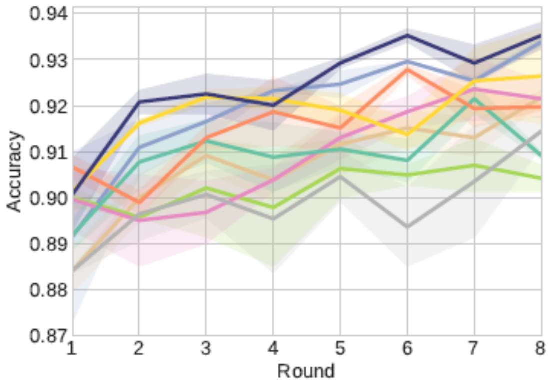
- Task: AL for text classification
- Model: RoBERTa-base
- Datasets:

Table 1: Dataset statistics.

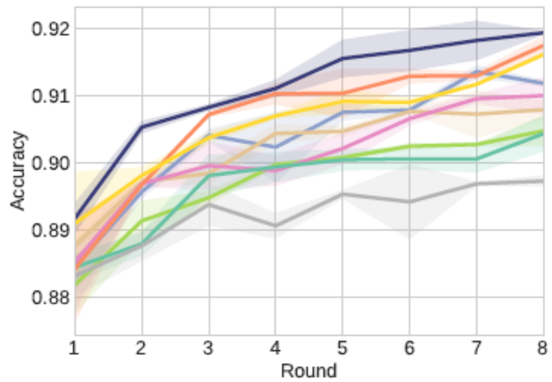
DATASET	LABEL TYPE	#TRAIN	#VAL	#TEST	#CLASSES
SST-2	Sentiment	40K	3K	1.8K	2
AGNEWS	News Topic	80K	3K	7.6K	4
PUBMED	Medical Abstract	100K	3K	30.1K	5
SNIPS	Intent	13K	0.7K	0.7K	7
STOV	Question	8.0K	1K	1K	10

- Eight baselines

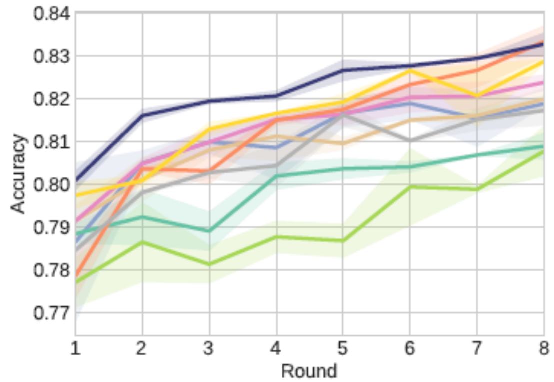
Results - Accuracy



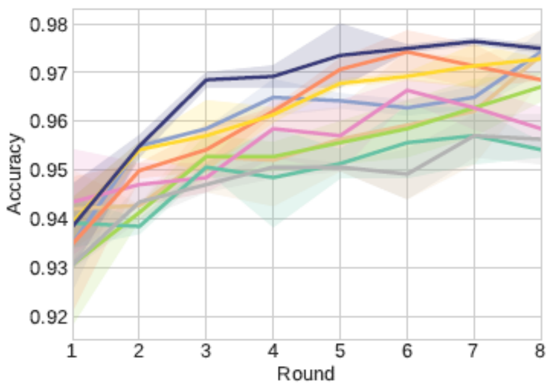
(a) SST-2



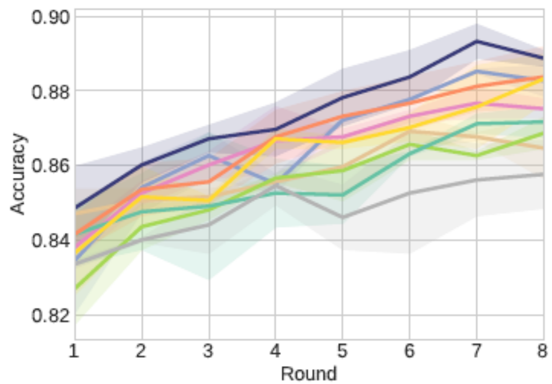
(b) AGNEWS



(c) PUBMED



(d) SNIPS



(e) STOV

(f) Legend

Results:
Error Rate

REAL: A Representative Error-Driven
Approach for
Active Learning

Results: Error Rate

$\varepsilon(Q)$

Error rate of the actively selected instances Q .

$\varepsilon(\mathcal{D}_u)$

Error rate of the whole unlabeled pool (as test set).

`lift`

$\varepsilon(Q)/\varepsilon(\mathcal{D}_u)$

ℓ_0

Average first step training loss for the the actively selected instances Q.

DATASET	METRIC	ENTROPY	PLM-KM	BADGE	CAL	ACTUNE	RANDOM	REAL
SST-2	$\varepsilon(Q)$	**0.4959**	0.1841	0.2308	0.4821	0.4334	0.1284	0.4739
	$\varepsilon(\mathcal{D}_u)$	**0.1194**	0.1251	0.1259	0.1215	0.1170	0.1338	0.1212
	lift	**4.1530**	1.4713	1.8325	3.9670	3.7055	0.9596	3.9113
	ℓ_0	0.6984	0.8100	**1.0538**	0.6915	0.8526	0.6660	0.9938
AGNEWS	$\varepsilon(Q)$	**0.6092**	0.1904	0.2246	0.5637	0.5325	0.1142	0.5537
	$\varepsilon(\mathcal{D}_u)$	0.1009	0.1039	0.1041	0.0995	0.0991	0.1115	**0.0959**
	lift	**6.0377**	1.8320	2.1576	5.6667	5.3730	1.0239	5.7737
	ℓ_0	1.2504	0.8597	0.9477	1.0926	1.3009	0.5707	**1.3636**
PUBMED	$\varepsilon(Q)$	**0.6701**	0.3164	0.3634	0.6103	0.6231	0.1987	0.6046
	$\varepsilon(\mathcal{D}_u)$	0.1943	0.1971	0.1928	0.1941	0.1907	0.1998	**0.1858**
	lift	**3.4487**	1.6048	1.8845	3.1452	3.2670	0.9943	3.2531
	ℓ_0	1.5117	1.3533	1.6009	1.2871	1.4494	1.0222	**1.7040**
SNIPS	$\varepsilon(Q)$	0.4107	0.1226	0.1120	**0.4237**	0.2963	0.0276	0.4002
	$\varepsilon(\mathcal{D}_u)$	0.0268	0.0337	0.0308	0.0280	0.0265	0.0393	**0.0231**
	lift	15.3183	3.6410	3.6338	15.1568	11.1895	0.7023	**17.2902**
	ℓ_0	**1.0176**	0.5209	0.5080	1.0470	0.9491	0.1842	0.9356
STOV	$\varepsilon(Q)$	**0.7328**	0.2536	0.3506	0.6904	0.6659	0.1307	0.7162
	$\varepsilon(\mathcal{D}_u)$	0.1048	0.1263	0.1209	0.1094	0.1101	0.1386	**0.1045**
	lift	**6.9934**	2.0079	2.8994	6.3114	6.0509	0.9435	6.8548
	ℓ_0	**2.1434**	1.0260	1.3874	2.0255	2.0062	0.6331	2.1131

Results: Error Rate

$\varepsilon(Q)$

Error rate of the actively selected instances Q .

$\varepsilon(\mathcal{D}_u)$

Error rate of the whole unlabeled pool (as test set).

`lift`

$\varepsilon(Q)/\varepsilon(\mathcal{D}_u)$

ℓ_0

Average first step training loss for the the actively selected instances Q.

DATASET	METRIC	ENTROPY	PLM-KM	BADGE	CAL	ACTUNE	RANDOM	REAL
SST-2	$\varepsilon(Q)$	**0.4959**	0.1841	0.2308	0.4821	0.4334	0.1284	0.4739
	$\varepsilon(\mathcal{D}_u)$	**0.1194**	0.1251	0.1259	0.1215	0.1170	0.1338	0.1212
	`lift`	**4.1530**	1.4713	1.8325	3.9670	3.7055	0.9596	3.9113
	ℓ_0	0.6984	0.8100	**1.0538**	0.6915	0.8526	0.6660	0.9938
AGNEWS	$\varepsilon(Q)$	**0.6092**	0.1904	0.2246	0.5637	0.5325	0.1142	0.5537
	$\varepsilon(\mathcal{D}_u)$	0.1009	0.1039	0.1041	0.0995	0.0991	0.1115	**0.0959**
	`lift`	**6.0377**	1.8320	2.1576	5.6667	5.3730	1.0239	5.7737
	ℓ_0	1.2504	0.8597	0.9477	1.0926	1.3009	0.5707	**1.3636**
PUBMED	$\varepsilon(Q)$	**0.6701**	0.3164	0.3634	0.6103	0.6231	0.1987	0.6046
	$\varepsilon(\mathcal{D}_u)$	0.1943	0.1971	0.1928	0.1941	0.1907	0.1998	**0.1858**
	`lift`	**3.4487**	1.6048	1.8845	3.1452	3.2670	0.9943	3.2531
	ℓ_0	1.5117	1.3533	1.6009	1.2871	1.4494	1.0222	**1.7040**
SNIPS	$\varepsilon(Q)$	0.4107	0.1226	0.1120	**0.4237**	0.2963	0.0276	0.4002
	$\varepsilon(\mathcal{D}_u)$	0.0268	0.0337	0.0308	0.0280	0.0265	0.0393	**0.0231**
	`lift`	15.3183	3.6410	3.6338	15.1568	11.1895	0.7023	**17.2902**
	ℓ_0	**1.0176**	0.5209	0.5080	1.0470	0.9491	0.1842	0.9356
STOV	$\varepsilon(Q)$	**0.7328**	0.2536	0.3506	0.6904	0.6659	0.1307	0.7162
	$\varepsilon(\mathcal{D}_u)$	0.1048	0.1263	0.1209	0.1094	0.1101	0.1386	**0.1045**
	`lift`	**6.9934**	2.0079	2.8994	6.3114	6.0509	0.9435	6.8548
	ℓ_0	**2.1434**	1.0260	1.3874	2.0255	2.0062	0.6331	2.1131

Results: Error Rate

$\varepsilon(Q)$

Error rate of the actively selected instances Q .

$\varepsilon(\mathcal{D}_u)$

Error rate of the whole unlabeled pool (as test set).

lift

$\varepsilon(Q)/\varepsilon(\mathcal{D}_u)$

ℓ_0

Average first step training loss for the the actively selected instances Q.

DATASET	METRIC	ENTROPY	PLM-KM	BADGE	CAL	ACTUNE	RANDOM	REAL
SST-2	$\varepsilon(Q)$	**0.4959**	0.1841	0.2308	0.4821	0.4334	0.1284	0.4739
	$\varepsilon(\mathcal{D}_u)$	**0.1194**	0.1251	0.1259	0.1215	0.1170	0.1338	0.1212
	lift	**4.1530**	1.4713	1.8325	3.9670	3.7055	0.9596	3.9113
	ℓ_0	0.6984	0.8100	**1.0538**	0.6915	0.8526	0.6660	0.9938
AGNEWS	$\varepsilon(Q)$	**0.6092**	0.1904	0.2246	0.5637	0.5325	0.1142	0.5537
	$\varepsilon(\mathcal{D}_u)$	0.1009	0.1039	0.1041	0.0995	0.0991	0.1115	**0.0959**
	lift	**6.0377**	1.8320	2.1576	5.6667	5.3730	1.0239	5.7737
	ℓ_0	1.2504	0.8597	0.9477	1.0926	1.3009	0.5707	**1.3636**
PUBMED	$\varepsilon(Q)$	**0.6701**	0.3164	0.3634	0.6103	0.6231	0.1987	0.6046
	$\varepsilon(\mathcal{D}_u)$	0.1943	0.1971	0.1928	0.1941	0.1907	0.1998	**0.1858**
	lift	**3.4487**	1.6048	1.8845	3.1452	3.2670	0.9943	3.2531
	ℓ_0	1.5117	1.3533	1.6009	1.2871	1.4494	1.0222	**1.7040**
SNIPS	$\varepsilon(Q)$	0.4107	0.1226	0.1120	**0.4237**	0.2963	0.0276	0.4002
	$\varepsilon(\mathcal{D}_u)$	0.0268	0.0337	0.0308	0.0280	0.0265	0.0393	**0.0231**
	lift	15.3183	3.6410	3.6338	15.1568	11.1895	0.7023	**17.2902**
	ℓ_0	**1.0176**	0.5209	0.5080	1.0470	0.9491	0.1842	0.9356
STOV	$\varepsilon(Q)$	**0.7328**	0.2536	0.3506	0.6904	0.6659	0.1307	0.7162
	$\varepsilon(\mathcal{D}_u)$	0.1048	0.1263	0.1209	0.1094	0.1101	0.1386	**0.1045**
	lift	**6.9934**	2.0079	2.8994	6.3114	6.0509	0.9435	6.8548
	ℓ_0	**2.1434**	1.0260	1.3874	2.0255	2.0062	0.6331	2.1131

Results: Error Rate

$\varepsilon(Q)$

Error rate of the actively selected instances Q .

$\varepsilon(\mathcal{D}_u)$

Error rate of the whole unlabeled pool (as test set).

`lift`

$\varepsilon(Q)/\varepsilon(\mathcal{D}_u)$

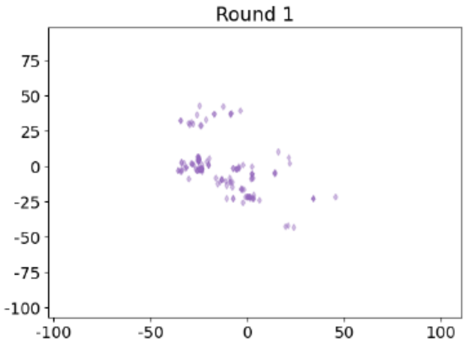
ℓ_0

Average first step training loss for the the actively selected instances Q. [3]

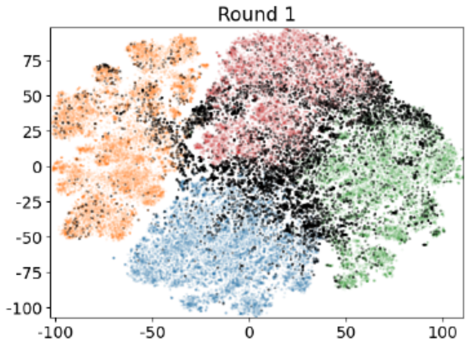
DATASET	METRIC	ENTROPY	PLM-KM	BADGE	CAL	ACTUNE	RANDOM	REAL
SST-2	$\varepsilon(Q)$	**0.4959**	0.1841	0.2308	0.4821	0.4334	0.1284	0.4739
	$\varepsilon(\mathcal{D}_u)$	**0.1194**	0.1251	0.1259	0.1215	0.1170	0.1338	0.1212
	lift	**4.1530**	1.4713	1.8325	3.9670	3.7055	0.9596	3.9113
	ℓ_0	0.6984	0.8100	**1.0538**	0.6915	0.8526	0.6660	0.9938
AGNEWS	$\varepsilon(Q)$	**0.6092**	0.1904	0.2246	0.5637	0.5325	0.1142	0.5537
	$\varepsilon(\mathcal{D}_u)$	0.1009	0.1039	0.1041	0.0995	0.0991	0.1115	**0.0959**
	lift	**6.0377**	1.8320	2.1576	5.6667	5.3730	1.0239	5.7737
	ℓ_0	1.2504	0.8597	0.9477	1.0926	1.3009	0.5707	**1.3636**
PUBMED	$\varepsilon(Q)$	**0.6701**	0.3164	0.3634	0.6103	0.6231	0.1987	0.6046
	$\varepsilon(\mathcal{D}_u)$	0.1943	0.1971	0.1928	0.1941	0.1907	0.1998	**0.1858**
	lift	**3.4487**	1.6048	1.8845	3.1452	3.2670	0.9943	3.2531
	ℓ_0	1.5117	1.3533	1.6009	1.2871	1.4494	1.0222	**1.7040**
SNIPS	$\varepsilon(Q)$	0.4107	0.1226	0.1120	**0.4237**	0.2963	0.0276	0.4002
	$\varepsilon(\mathcal{D}_u)$	0.0268	0.0337	0.0308	0.0280	0.0265	0.0393	**0.0231**
	lift	15.3183	3.6410	3.6338	15.1568	11.1895	0.7023	**17.2902**
	ℓ_0	**1.0176**	0.5209	0.5080	1.0470	0.9491	0.1842	0.9356
STOV	$\varepsilon(Q)$	**0.7328**	0.2536	0.3506	0.6904	0.6659	0.1307	0.7162
	$\varepsilon(\mathcal{D}_u)$	0.1048	0.1263	0.1209	0.1094	0.1101	0.1386	**0.1045**
	lift	**6.9934**	2.0079	2.8994	6.3114	6.0509	0.9435	6.8548
	ℓ_0	**2.1434**	1.0260	1.3874	2.0255	2.0062	0.6331	2.1131

[3] Yoo et al., Learning Loss for Active Learning, CVPR'2019

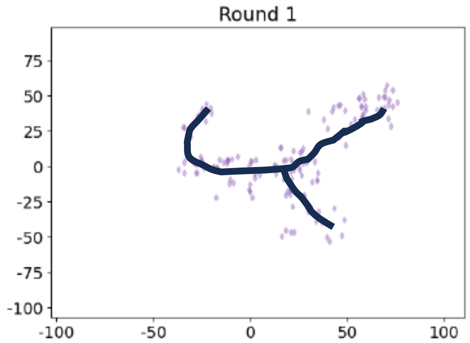
Results – Representative Errors



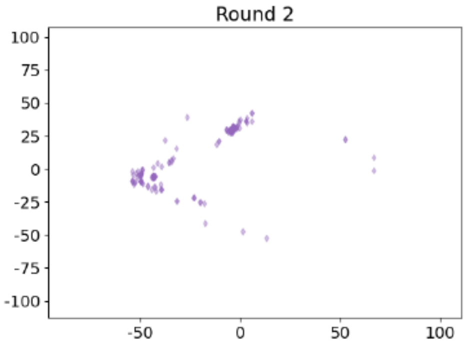
(a) ENTROPY



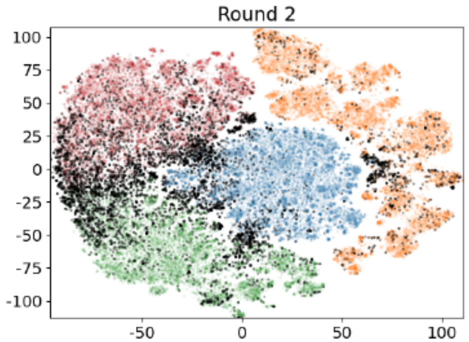
(b) Errors (in black)



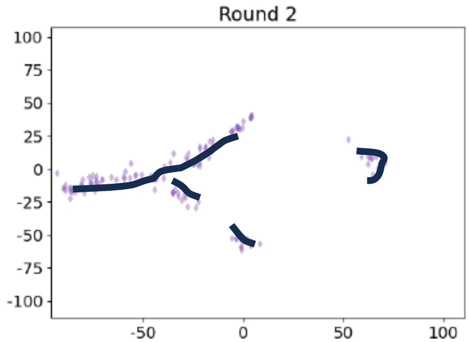
(c) REAL



(d) ENTROPY

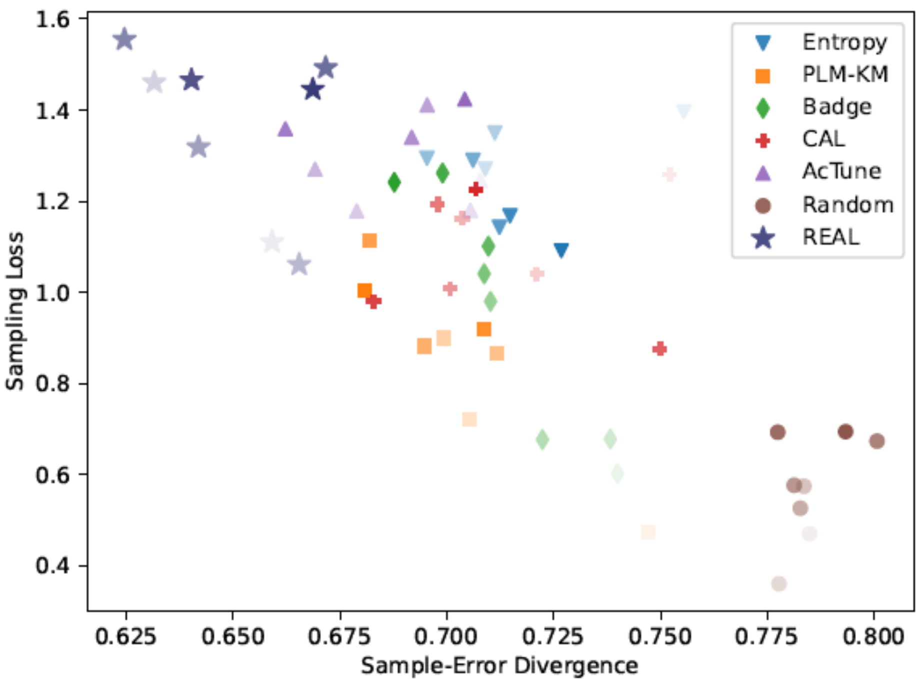


(e) Errors (in black)

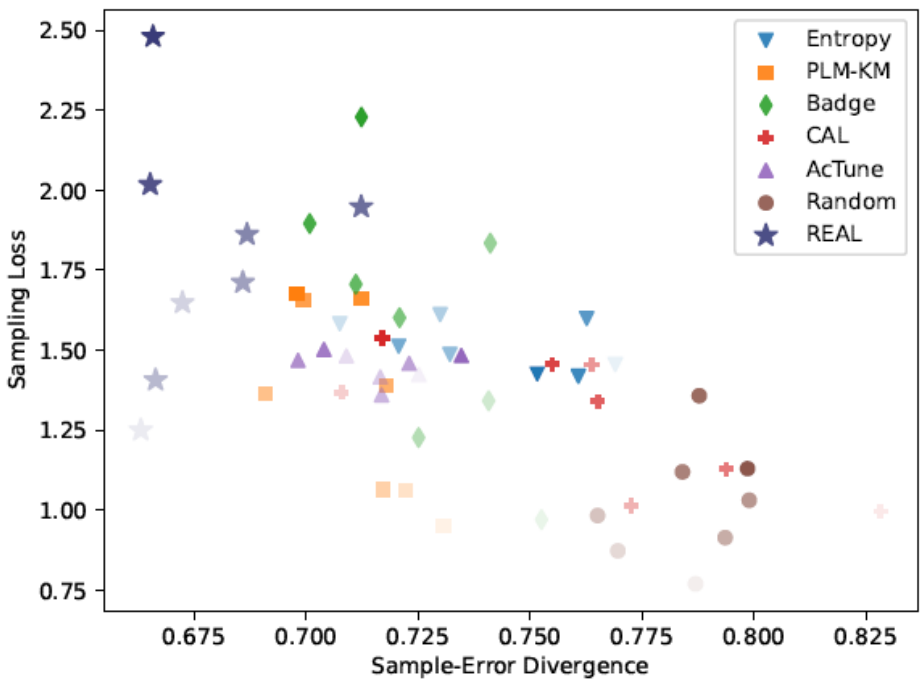


(f) REAL

Results – Representative Errors



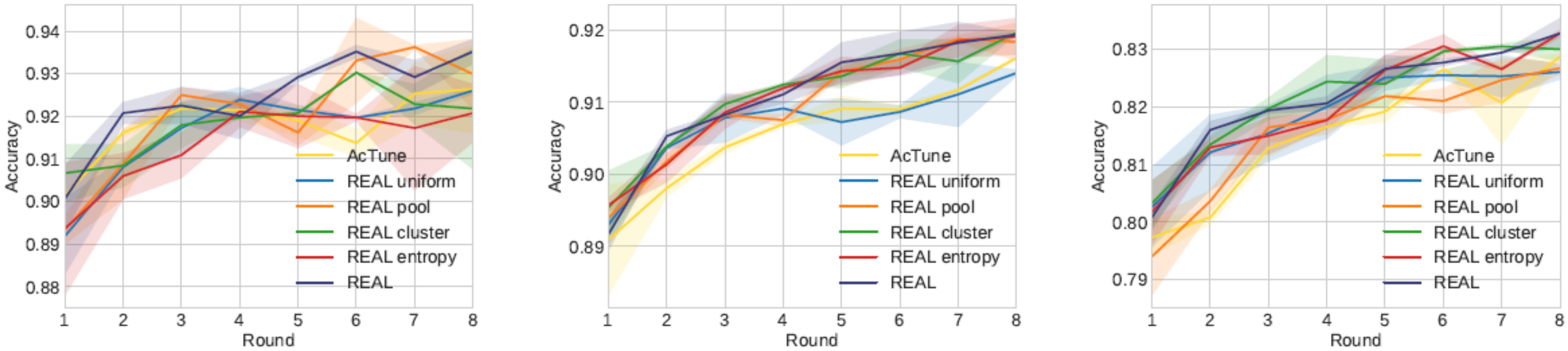
(a) AGNEWS



(b) PUBMED

Ablation Study

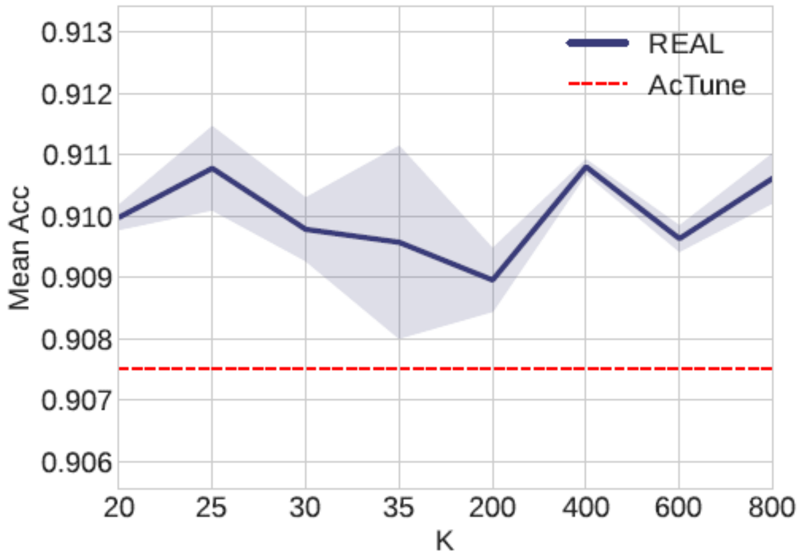
- Most variants of REAL still performs well



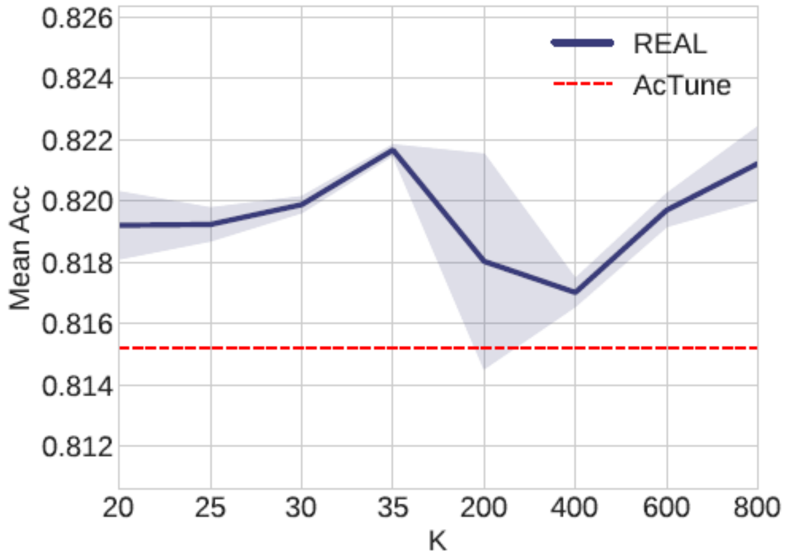
(a) SST-2 (b) AGNEWS (c) PUBMED

Hyperparameter

- Mean acc under a wide range of #clusters

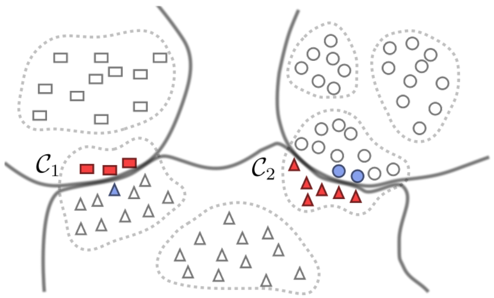


(a) AGNEWS



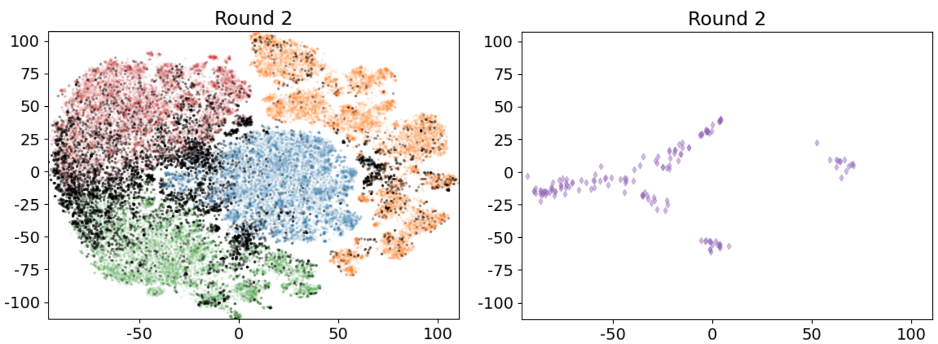
(b) PUBMED

Takeaways



REAL: a new AL sampling algorithm for Representative Errors

- Key: adaptive budget allocation



Most unlabeled errors lie around the decision boundary

- Finding those errors for labeling can improve AL

Thank you for your attention!

Q & A

REAL: A Representative Error-Driven Approach for Active Learning

Cheng Chen[1,2], Yong Wang[2], Lizi Liao[2], Yueguo Chen[1], Xiaoyong Du[1]

Code & data: https://github.com/withchencheng/ECML_PKDD_23_Real
Contact me: chchen@ruc.edu.cn