Prefer to Classify: Improving Text Classifiers via Auxiliary Preference Learning

Jaehyung Kim[1] Jinwoo Shin[1] Deogyeop Kang[2]

[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]University of Minnesota (UMN)

Importance of NLP Benchmarks

- Success of NLP systems has been driven by **large human-annotated benchmarks**
  - They guide the researchers in a right direction to develop methods
  - E.g., SQuAD (QA), GLUE (language understanding), and BIG-bench (large language models)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

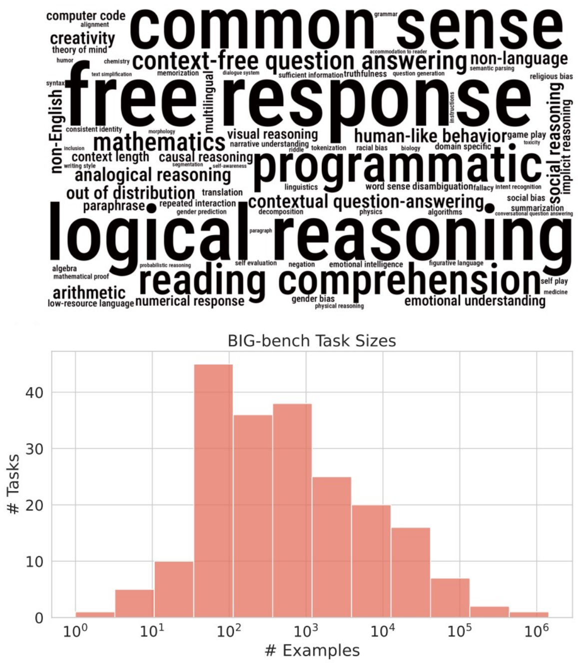
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Example of SQuAD (100k+) [Rajpurkar et al. 2016]

Corpus	|Train|	|Test|	Task
CoLA	8.5k	**1k**	acceptability
SST-2	67k	1.8k	sentiment
MRPC	3.7k	1.7k	paraphrase
STS-B	7k	1.4k	sentence similarity
QQP	364k	**391k**	paraphrase
MNLI	393k	**20k**	NLI
QNLI	105k	5.4k	QA/NLI
RTE	2.5k	3k	NLI
WNLI	634	**146**	coreference/NLI

Summary of GLUE [Wang et al. 2019]

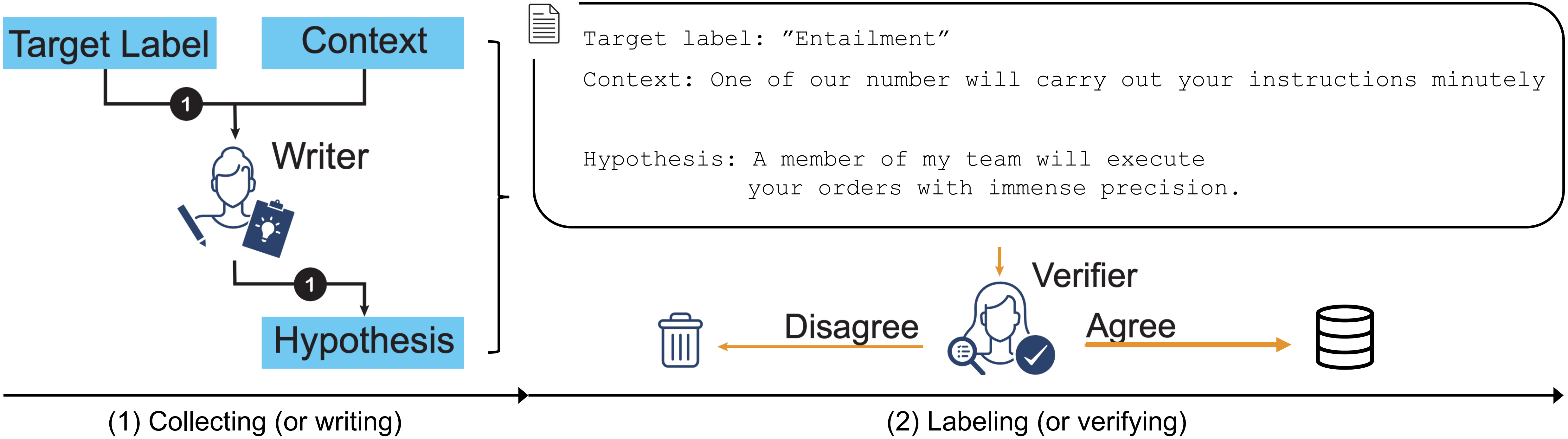


Diversity/scale of BIG-bench [Srivastava et al. 2022]

[Rajpurkar et al. 2016] SQuAD: 100,000+ Questions for Machine Comprehension of Text, EMNLP 2016
[Wang et al. 2019] GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding, ICLR 2019
[Srivastava et al. 2022] Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models, arXiv:2206.04615

Construction of NLP Benchmarks

- These benchmarks are usually constructed by following steps
  1. **Collecting (or writing)** the relevant input texts
  2. **Labeling** input texts **(or verifying)** by human annotators

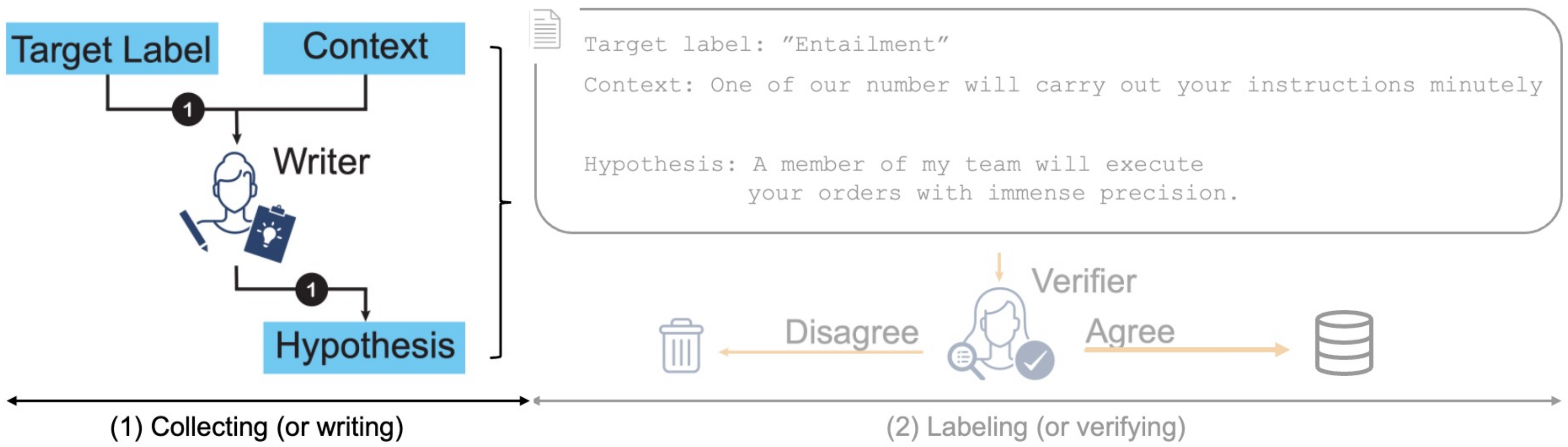


Example of procedure for constructing benchmark for NLI task [Nie et al. 2019]

[Nie et al. 2019] Adversarial NLI: A New Benchmark for Natural Language Understanding, ACL 2019

Cost for Constructing NLP Benchmarks

- These benchmarks are usually constructed by following steps
  1. **Collecting (or writing)** the relevant input texts ⟶ **more costly and cumbersome**
     - E.g., distribution shift or spurious patterns of input make model suffer being generalized [Gururangan et al. 2018; Karamcheti et al. 2021]
     - Hence, much higher cost is often paid to the collection process to keep the quality [Kaushik et al. 2020]
  2. **Labeling** input texts **(or verifying)** by human annotators



[Gururangan et al. 2018] Annotation Artifacts in Natural Language Inference Data, NAACL 2018
[Karamcheti et al. 2021] Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering, ACL 2021
[Kaushik et al. 2020] Learning the Difference that Makes a Difference with Counterfactually-augmented Data, ICLR 2020

Complementary Way to Annotate Existing Benchmarks

- Hence, it is preferable to pay additional human cost to **auxiliary annotation**
  - E.g., improving label quality with more annotators [Nie et al. 2020]
  - or obtaining finer task information with new label space [Williams et al. 2020]

Context	Hypothesis	Old Labels	New Labels	Source	Type
		majority and individual labels			
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** E E E N N	**Entailment** E(50) N(50)	SNLI	Low agreements
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** E E N N E	**Neutral** N(93) E(7)	SNLI	Majority changed
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** C N N C C	**Contradiction** C(56) N(44)	MNLI	Low agreements
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** N E N N E	**Entailment** E(94) N(5) C(1)	MNLI	Majority changed

Analysis of existing NLI datasets with more annotations [Nie et al. 2020]

Dataset	Subset	Numerical	Basic	Reference	Tricky	Reasoning	Error
A1	All	40.8	31.4	24.5	29.5	58.4	3.3
	C	18.6	8.2	7.8	13.7	11.9	0.7
	N	7.0	9.8	7.1	6.4	31.3	1.0
	E	15.2	13.4	9.6	9.4	15.2	1.6
A2	All	38.5	41.2	29.4	29.1	62.7	2.5
	C	15.6	11.8	10.2	13.6	15.5	0.3
	N	8.1	12.8	9.1	7.4	30.0	1.4
	E	14.8	16.6	10.1	8.1	17.2	0.8
A3	All	20.3	50.2	27.5	25.6	63.9	2.2
	C	8.7	17.2	8.6	12.7	14.9	0.3
	N	4.9	13.1	8.2	4.6	30.1	1.0
	E	6.7	19.9	10.7	8.3	18.9	0.8

Analysis of ANLI with fine-grained annotation [Williams et al. 2020]

[Nie et al. 2020] What Can We Learn from Collective Human Opinions on Natural Language Inference Data?, EMNLP 2020
[Williams et al. 2020] ANLIzing the Adversarial Natural Language Inference Dataset, arXiv:2010.12729

Complementary Way to Annotate Existing Benchmarks

- Hence, it is preferable to pay additional human cost to **auxiliary annotation**
  - E.g., improving label quality with more annotators [Nie et al. 2020]
  - or obtaining finer task information with new label space [Williams et al. 2020]

Context	Hypothesis	Old Labels	New Labels majority and individual labels	Source	Type
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** E E E N N	**Entailment** $E^{(50)}$ $N^{(50)}$	SNLI	Low agreements
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** E E N N E	**Neutral** $N^{(93)}$ $E^{(7)}$	SNLI	Majority changed
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** C N N C C	**Contradiction** $C^{(56)}$ $N^{(44)}$	MNLI	Low agreements
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** N E N N E	**Entailment** $E^{(94)}$ $N^{(5)}$ $C^{(1)}$	MNLI	Majority changed

Dataset	Subset	Numerical	Basic	Reference	Tricky	Reasoning	Error
A1	All	40.8	31.4	24.5	29.5	58.4	3.3
	C	18.6	8.2	7.8	13.7	11.9	0.7
	N	7.0	9.8	7.1	6.4	31.3	1.0
	E	15.2	13.4	9.6	9.4	15.2	1.6
A2	All	38.5	41.2	29.4	29.1	62.7	2.5
	C	15.6	11.8	10.2	13.6	15.5	0.3
	N	8.1	12.8	9.1	7.4	30.0	1.4
	E	14.8	16.6	10.1	8.1	17.2	0.8
A3	All	20.3	50.2	27.5	25.6	63.9	2.2
	C	8.7	17.2	8.6	12.7	14.9	0.3
	N	4.9	13.1	8.2	4.6	30.1	1.0
	E	6.7	19.9	10.7	8.3	18.9	0.8

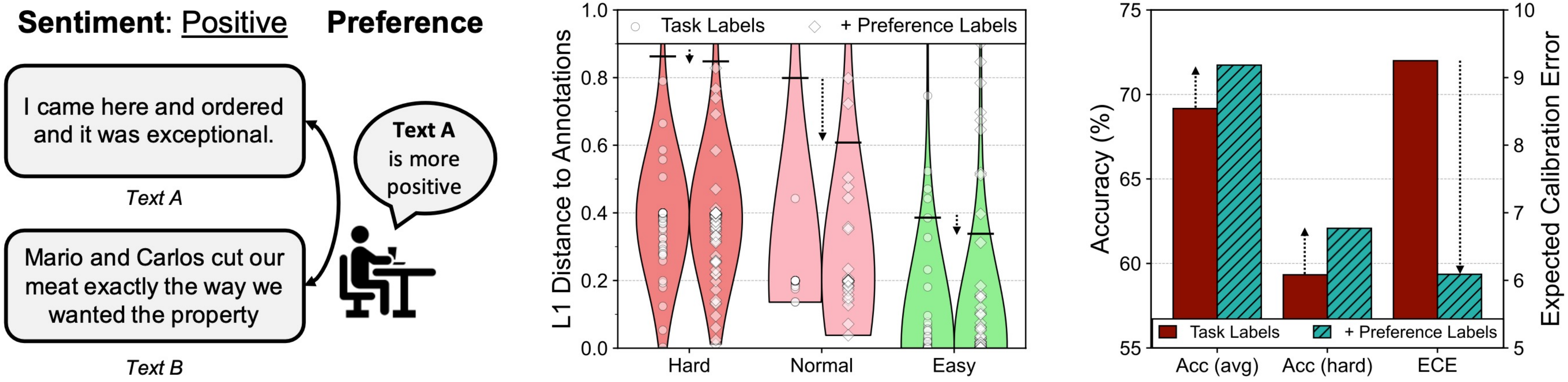
Can we find a **new alternative way to better exploit** existing benchmarks (input texts and task labels) via **auxiliary annotation**?

Especially, for text classification

Task-specific Preference as Auxiliary Annotation

- Idea: using **task-specific preference** between input texts as auxiliary annotation
  - To improve the text classification system upon existing task annotations
  - Auxiliary preference learning provides additional informative training signal to model
    - By relatively ordering a pair of two texts and better calibrating them w.r.t task through "pair-wise" comparison

v.s. "instance-wise" task annotation

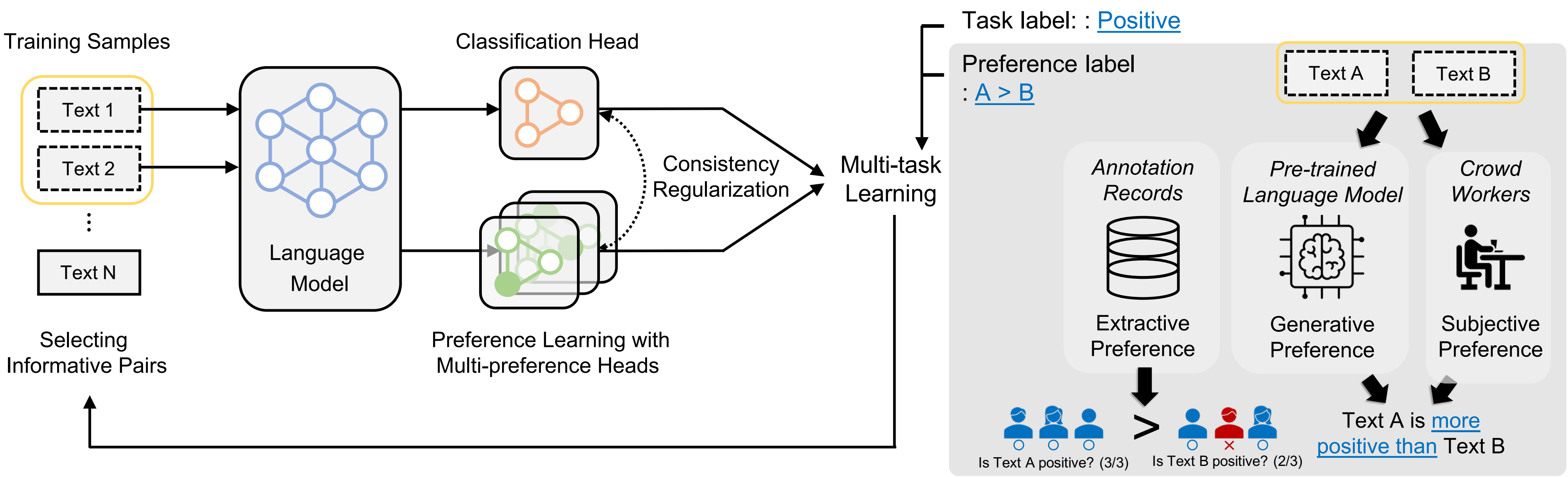


(a) Pair-wise preference signals (b) Alignment to human annotations (c) Improved text classification

Concept of auxiliary preference learning and its empirical advantages

Prefer to Classify (P2C)

- Specifically, we propose following components for auxiliary preference learning
  - **Three different types of preference labels** in practical scenario
    - Using large language model (generative), data annotation records (extractive), or crowd workers (subjective)
  - **Novel multi-task learning** framework with task and preference labels: prefer-to-classify (P2C)



Multi-task learning with task and preference labels

Preference label with three different ways

Visual illustration of the proposed auxiliary preference learning for improving text classifier

Different Types of Preference Labels

- **3 different types of preference** labels to apply auxiliary preference learning via P2C
  - **Generative preference** from large language models, e.g., GPT-3 [Brown et al. 2020]
    - Good quality from strong zero/few-shot generalization capability of LM, low cost, and **easy to access**

```
Read given two sentences A and B, and pick a more {labels_ab[idx]} sentence:.
    Sentence A: {sentences_a[idx]}
    Sentence B: {sentences_b[idx]}
    Choices: [Sentence A, Sentence B, No Preference], Answer:
```

Prompt design to collect generative preference labels from GPT-3 [Brown et al. 2020]

[Brown et al. 2020] Language Models are Few-shot Learners, NeurIPS 2020

Different Types of Preference Labels

- **3 different types of preference** labels to apply auxiliary preference learning via P2C
  - **Extractive preference** from data annotation records
    - If one sample has higher voting than the other sample as specific label, then it is assumed to be more preferred
    - **Zero cost** with good quality by better exploiting information within task annotions, but often hard to access

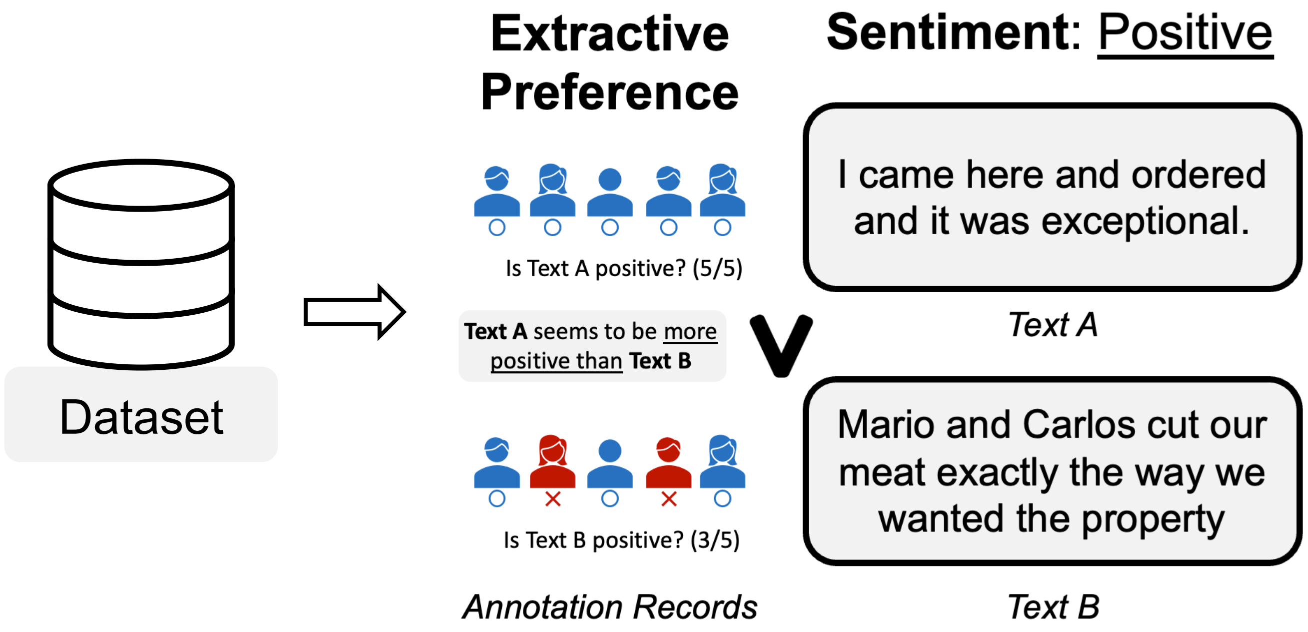
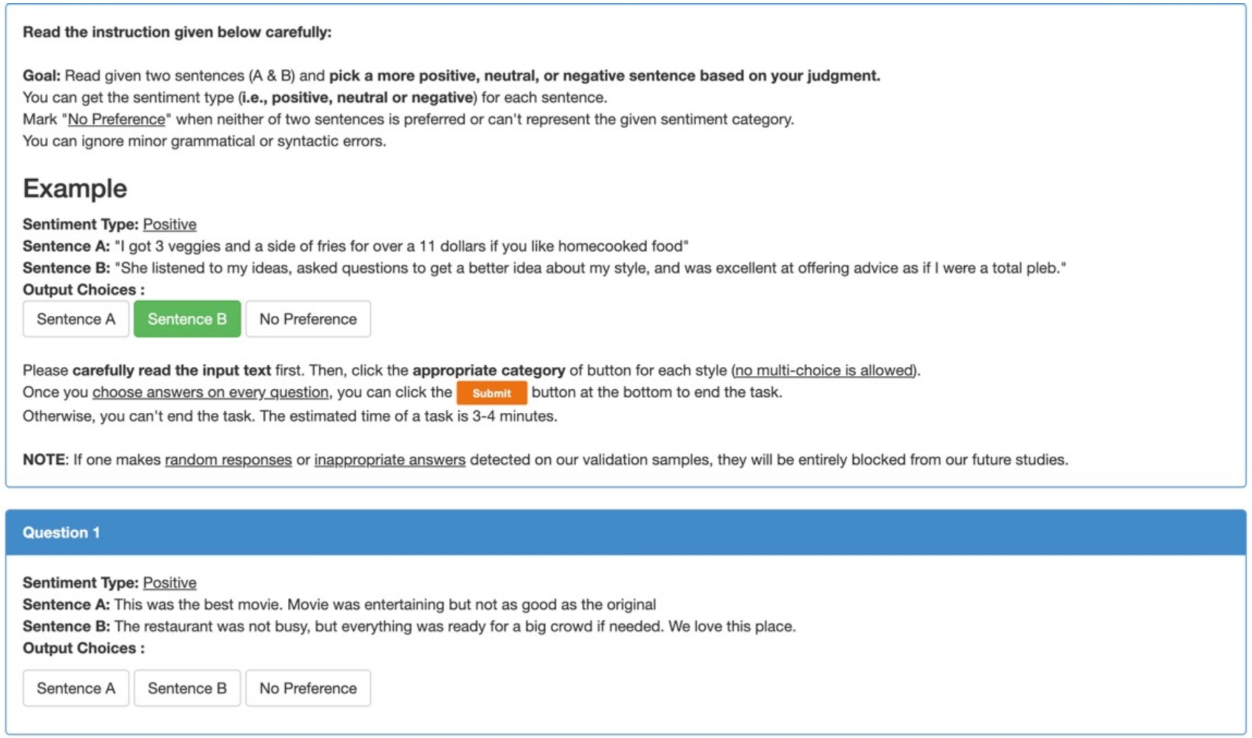


Illustration of extractive preference from existing data annotation records

Different Types of Preference Labels

- **3 different types of preference** labels to apply auxiliary preference learning via P2C
  - **Subjective preference** from crowd workers
    - Obtained by directly asking the to humans, e.g., “*which sentence is more positive?*”
    - **Most accurate**, but it requires high cost and hence hard to access



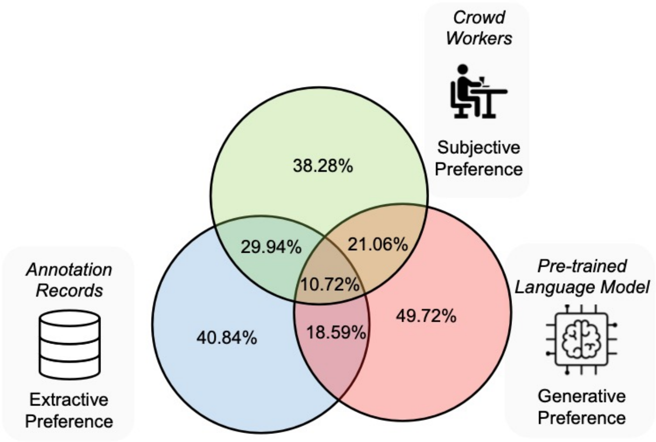
Used interface to collect subjective preference labels from crowd workers via AMT

Different Types of Preference Labels

- **3 different types of preference** labels to apply auxiliary preference learning via P2C
  - **Generative** / **Extractive** / **Subjective** preference labels
    - Accuracy: subjective > extractive ~= generative
    - Cost: extractive > generative >> subjective (e.g., 1.6$ for 10 subjective labels, while 8.0$ for 5,000 generative labels)
    - Accessibility: generative > extractive > subjective

A: We enjoyed our first and last meal in Toronto at Bombay Palace, and I can't think of a better way to book our journey.	**B**: So glad I finally tried this place because if confirmed my suspicions about that critic who rated it a 10.
Sentiment: Positive, Generative Preference: **A** $\succ$ **B**, Extractive Preference: **B** $\succ$ **A**, Subjective Preference: **No preference**	
A: The buffalo chicken was not good, but very costly.	**B**: There was so much stuff from all over that I had to leave to find an ATM for more cash to pay for it all.
Sentiment: Negative, Generative Preference: **A** $\succ$ **B**, Extractive Preference: **B** $\succ$ **A**, Subjective Preference: **B** $\succ$ **A**	
A: The hotel offered complimentary breakfast.	**B**: My friends had a full acrylic and the other had a fill. It looked so good.
Sentiment: Positive, Generative Preference: **A** $\succ$ **B**, Extractive Preference: **A** $\succ$ **B**, Subjective Preference: **A** $\succ$ **B**	

Examples of the collected preference labels on same pair of sentences from DynaSent-R2 [Potts et al. 2021]



Overlap between preferences

[Potts et al. 2021] Dynasent: A Dynamic Benchmark for Sentiment Analysis, ACL 2021

Prefer to Classify (P2C): Multi-task Learning

- Classifier is trained using task labels and preference labels jointly with
  - **Diverse multi-preference heads** for better preference modeling
    - For preference predictor, we add preference prediction head $W_{\texttt{pref}}$ on classifier $g_\phi(\mathbf{x})$ (e.g., BERT)

$$P_\psi[\mathbf{x}^1 \succ \mathbf{x}^0; y_{\texttt{task}}] = \frac{\exp\left(h_\psi(\mathbf{x}^1, y_{\texttt{task}})\right)}{\sum_{i\in\{0,1\}} \exp\left(h_\psi(\mathbf{x}^i, y_{\texttt{task}})\right)} \qquad h_\psi(\mathbf{x}, y_{\texttt{task}}) = W_{\texttt{pref}} \circ [g_\phi(\mathbf{x}); y_{\texttt{task}}]$$

$$\mathcal{L}_{\texttt{pref}} = -\mathop{\mathbb{E}}_{\substack{(\mathbf{x}^0,\mathbf{x}^1,y_{\texttt{task}},y_{\texttt{pref}})\\ \sim\mathcal{D}}}\Big[y_{\texttt{pref}} \log P_\psi[\mathbf{x}^1 \succ \mathbf{x}^0; y_{\texttt{task}}] + (1 - y_{\texttt{pref}}) \log P_\psi[\mathbf{x}^0 \succ \mathbf{x}^1; y_{\texttt{task}}]\Big]$$

    - Then, we introduce multiple preference heads $\{W_{\texttt{pref}}^{(t)}\}_{t=1}^T$ and maximize KL divergence between their prediction

$$\mathcal{L}_{\texttt{div}} = \frac{-1}{T-1} \sum_{j=1, j\neq i}^{T} D_{\text{KL}}\left(P_{\psi^{(i)}}(\mathbf{x}^1, \mathbf{x}^0; y_{\texttt{task}}) || P_{\psi^{(j)}}(\mathbf{x}^1, \mathbf{x}^0; y_{\texttt{task}})\right)$$

    - Overall multi-task learning objective

$$\mathcal{L}_{\texttt{multi}} = \mathcal{L}_{\texttt{task}} + \mathcal{L}_{\texttt{pref}}^{\text{all}} + \lambda_{\texttt{div}} \mathcal{L}_{\texttt{div}} \qquad \mathcal{L}_{\texttt{pref}}^{\text{all}} = \sum_{t=1}^{T} \mathcal{L}_{\texttt{pref}}^{\psi^{(t)}}$$

Prefer to Classify (P2C): Multi-task Learning

- Classifier is trained using task labels and preference labels jointly with
  - **Diverse multi-preference heads** for better preference modeling
  - **Consistency regularization** between task and preference learning
    - To explicitly impose the intuition: "*preferred instance should have a higher confidence*"
    - To this end, applying following consistency loss

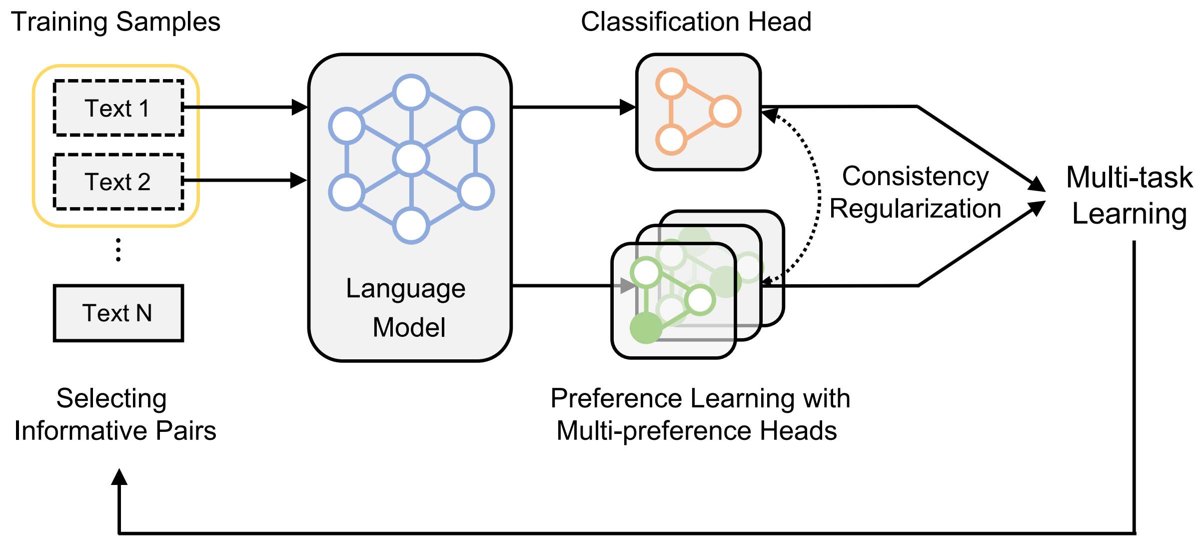
$$\mathcal{L}_{\texttt{cons}} = y_{\texttt{pref}} \max\{0, p_y(\mathbf{x}^1) - p_y(\mathbf{x}^0)\} + (1 - y_{\texttt{pref}}) \max\{0, p_y(\mathbf{x}^0) - p_y(\mathbf{x}^1)\}$$

    - Overall, our training loss is as follow:

$$\mathcal{L}_{\texttt{train}} = \mathcal{L}_{\texttt{multi}} + \lambda_{\texttt{cons}} \mathcal{L}_{\texttt{cons}}$$

Prefer to Classify (P2C): Multi-task Learning

- Classifier is trained using task labels and preference labels jointly with
  - **Diverse multi-preference heads** for better preference modeling
  - **Consistency regularization** between task and preference learning
  - **Selecting informative pairs** of input texts
    - *Disagreement-based* sampling: selecting pairs with high variance across multiple preference predictors $\{h_{\psi^{(i)}}\}_{t=1}^{T}$
    - *Inconsistency-based* sampling: selecting pairs with high consistency loss $\mathcal{L}_{\mathtt{cons}}$



Experiments

- Text classification with **generative preference**
  - P2C is consistently effective in improving the performance (accuracy and calibration)
    - bAcc: balanced accuracy as datasets have imbalanced distribution, wAcc: worst-group accuracy (minority)
  - P2C also outperforms GPT-3 baselines ⟶ Not just distilling "instance-wise" knowledge of GPT-3

Method	CoLA		SMS Spam		Hate Speech		Emotion	
	Mcc(↑)	ECE(↓)	bAcc(↑) / wAcc(↑)	ECE(↓)	bAcc(↑) / wAcc(↑)	ECE(↓)	bAcc(↑) / wAcc(↑)	ECE(↓)
Vanilla	$63.7_{\pm1.0}$	$3.6_{\pm1.6}$	$96.9_{\pm0.3}$ / $95.1_{\pm1.5}$	$1.3_{\pm0.3}$	$81.1_{\pm1.8}$ / $69.9_{\pm4.6}$	$5.1_{\pm1.0}$	$88.6_{\pm2.3}$ / $76.1_{\pm7.8}$	$4.0_{\pm1.1}$
Label Smoothing	$63.9_{\pm0.3}$	$4.6_{\pm1.2}$	$96.9_{\pm0.8}$ / $94.0_{\pm1.5}$	$1.1_{\pm0.3}$	$81.5_{\pm0.9}$ / $71.3_{\pm3.2}$	$6.6_{\pm1.0}$	$89.8_{\pm0.8}$ / $76.9_{\pm6.6}$	$4.0_{\pm0.9}$
Max Entropy	$64.1_{\pm0.3}$	$4.5_{\pm0.4}$	$96.9_{\pm1.1}$ / $94.7_{\pm1.6}$	$1.2_{\pm0.3}$	$81.6_{\pm1.8}$ / $70.5_{\pm4.2}$	$4.3_{\pm0.7}$	$89.1_{\pm1.1}$ / $73.1_{\pm2.5}$	**$3.6_{\pm0.9}$**
CS-KD	$64.5_{\pm1.4}$	$4.1_{\pm1.1}$	$96.8_{\pm0.9}$ / $94.0_{\pm2.4}$	**$1.1_{\pm0.2}$**	$81.4_{\pm2.6}$ / $69.6_{\pm5.1}$	$5.3_{\pm1.8}$	$89.4_{\pm1.6}$ / $74.0_{\pm6.8}$	$4.1_{\pm0.2}$
GPT-3 (0-shot)	60.4	-	90.3 / 84.3	-	68.7 / 41.6	-	50.2 / 23.3	-
GPT-3 (5-shot)	$58.5_{\pm0.4}$	-	$92.2_{\pm0.5}$ / $88.5_{\pm0.7}$	-	$78.5_{\pm2.0}$ / $70.3_{\pm3.6}$	-	$46.6_{\pm0.6}$ / $30.3_{\pm2.6}$	-
GPT-3 (20-shot)	$58.3_{\pm1.4}$	-	$95.8_{\pm0.4}$ / $94.4_{\pm0.7}$	-	$77.8_{\pm0.5}$ / $69.0_{\pm1.5}$	-	$47.5_{\pm1.0}$ / $30.8_{\pm4.5}$	-
P2C (Ours)	**$65.4_{\pm1.0}$**	**$2.8_{\pm1.1}$**	**$97.4_{\pm0.4}$ / $95.2_{\pm1.0}$**	**$1.1_{\pm0.3}$**	**$82.4_{\pm1.3}$ / $73.6_{\pm4.5}$**	**$4.0_{\pm0.3}$**	**$90.7_{\pm0.7}$ / $81.7_{\pm4.7}$**	**$3.6_{\pm0.8}$**

11.55% relative test error reduction compared to *Vanilla*

Test accuracy of fine-tuned RoBERTa-base classifiers

Experiments

- Text classification with **extractive preference (Free!)**
  - P2C even outperforms the strong baselines for learning with annotation records

Method	Offensive	Polite-Wiki	Polite-SE	MNLI	DynaSent-R1	DynaSent-R2
Vanilla	75.88±0.72	89.35±1.53	70.00±1.49	81.92±0.70	80.43±0.30	71.23±1.05
Soft-labeling	76.08±1.44	89.57±1.76	70.35±1.68	82.67±0.50	81.10±0.33	72.15±1.59
Margin Loss	76.67±1.18	88.51±0.93	70.51±1.16	81.41±0.63	80.42±0.23	69.27±0.98
Filtering	76.13±1.18	89.50±0.87	68.28±2.43	82.13±0.67	80.38±0.34	69.86±0.78
Weighting	76.17±1.18	89.65±1.46	68.38±1.67	82.48±0.49	80.21±0.41	71.81±1.12
Multi-annotator	76.50±1.98	89.88±1.82	69.39±2.84	82.61±0.70	81.14±0.55	71.97±1.25
CS-KD	75.75±0.66	89.65±1.84	70.10±1.29	82.32±0.23	80.63±0.27	71.81±0.67
P2C (Ours)	**77.81**±0.21	**91.06**±0.64	**71.21**±0.93	**83.15**±0.29	**81.50**±0.39	**73.06**±0.31

7.59% / 4.27% relative test error reduction compared to *Vanilla* / *Best,* respectively

Test accuracy of fine-tuned RoBERTa-base classifiers

Experiments

- Comparison between different annotation methods
  - **Setup**: Given the existing datasets, adding the same number of annotations but different ways
  - **Results**
    - Overall, preference labels are effective for hard samples (i.e., high disagreement) along with strong calibration effects
    - Subjective preference labels are the most effective for improving accuracy and calibration

Method	$N_{\texttt{task}}$	$N_{\texttt{pref}}$	$\text{Acc}_{\texttt{avg}}(\uparrow)$	$\text{Acc}_{\texttt{hard}}$ / $\text{Acc}_{\texttt{easy}}(\uparrow)$	ECE($\downarrow$)	$\text{d}_{\texttt{hard}}$ / $\text{d}_{\texttt{easy}}(\downarrow)$
Vanilla	7.5k	-	$69.03_{\pm 1.29}$	$59.33_{\pm 2.57}$ / $80.00_{\pm 1.22}$	$9.25_{\pm 1.39}$	$0.856_{\pm 0.01}$ / $0.405_{\pm 0.03}$
Task Labels	12.5k	-	$71.17_{\pm 1.35}$	$57.86_{\pm 2.31}$ / $\mathbf{84.21}_{\pm 1.05}$	$9.19_{\pm 1.36}$	$0.878_{\pm 0.04}$ / $\mathbf{0.327}_{\pm 0.02}$
Generative Preference	7.5k	5k	$\underline{71.46}_{\pm 1.16}$	$\underline{61.77}_{\pm 0.94}$ / $82.28_{\pm 1.01}$	$\underline{6.64}_{\pm 0.79}$	$0.850_{\pm 0.02}$ / $0.361_{\pm 0.02}$
Extractive Preference	7.5k	5k	$71.36_{\pm 1.19}$	$61.16_{\pm 1.91}$ / $\underline{83.11}_{\pm 1.78}$	$6.75_{\pm 0.78}$	$\underline{0.847}_{\pm 0.03}$ / $\underline{0.351}_{\pm 0.03}$
Subjective Preference	7.5k	5k	$\mathbf{71.74}_{\pm 1.04}$	$\mathbf{62.08}_{\pm 0.94}$ / $83.01_{\pm 1.27}$	$\mathbf{6.09}_{\pm 0.31}$	$\mathbf{0.828}_{\pm 0.02}$ / $0.356_{\pm 0.02}$

Test accuracy of fine-tuned RoBERTa-base classifiers on DynaSent-R2

Summary

- We introduce **preference label** as new auxiliary annotation to improve benchmark
  - It provides additional informative training signal to model via "*pair-wise*" comparison
  - We propose an effective multi-task learning framework, coined *prefer-to-classify (P2C)*
  - We provide *three different ways* to obtain preference labels (generative/extractive/subjective)
- P2C shows consistent improvements on various NLP benchmarks
  - Improved test accuracy with better calibration
- P2C suggests new way to evolve benchmark along with recent advance of LM

For more details and results,
please see our paper and code

arXiv: 2306.04925

Github

Thank you for attention 😄