Towards Improving Selective Prediction Ability of NLP Systems

Neeraj Varshney, Swaroop Mishra, Chitta Baral
Arizona State University, USA

Selective Prediction

Selective Prediction allows a system to abstain from answering when its prediction is likely to be incorrect.

A Selective prediction system comprises of:
- A **predictor** function (f) that gives the model's prediction
- A **selector** function (g) that determines if the system should output the prediction made by the predictor.

Usually, '*g*' comprises of a confidence estimator '$\tilde{g}$' that indicates *f's* prediction confidence and a threshold '*th*' that controls the abstention level:

$$(f,g)(x) = \begin{cases} f(x), & \text{if g(x) = 1} \\ Abstain, & \text{if g(x) = 0} \end{cases} \qquad g(x) = \mathbb{1}[\tilde{g}(x)) > th]$$

A selective prediction system makes trade-offs between **coverage** and **risk**. For a dataset D, **coverage** at a threshold th is defined as the fraction of total instances answered by the system (where $\tilde{g} > th$) and **risk** is the error on the answered instances:

$$coverage_{th} = \frac{\sum_{x_i \in D} \mathbb{1}[\tilde{g}(x_i)) > th]}{|D|}$$

$$risk_{th} = \frac{\sum_{x_i \in D} \mathbb{1}[\tilde{g}(x_i)) > th] l_i}{\sum_{x_i \in D} \mathbb{1}[\tilde{g}(x_i)) > th]}$$

- With decrease in threshold, coverage will increase, but the risk will usually also increase.
- The overall selective prediction performance is measured by the area under Risk-Coverage curve (AUC).
- **Lower the AUC, the better the selective prediction system as it represents lower average risk across all thresholds.**

Proposed Method

- We leverage a held-out dataset and annotate it's instances such that the annotation score reflects the likelihood for the model's prediction to be correct.
- Then, we train a calibrator using this annotated held-out dataset and use it as the confidence estimator.

Annotating held-out instances:
- Annotation score is computed using maximum softmax probability (maxProb) of the model's prediction and difficulty score (d or 1 - s) of the instance.
- We demonstrate that maxProb is positively correlated while difficulty score is negatively correlated with the predictive correctness and explore three ways of computing this score:

$$AS_1 = \begin{cases} 0.5 + \frac{maxProb}{2}, & \text{if correct} \\ 0.5 - \frac{maxProb}{2}, & \text{otherwise} \end{cases}$$

$$AS_2 = \begin{cases} 0.5 + \frac{s_i}{2}, & \text{if correct} \\ 0.5 - \frac{s_i}{2}, & \text{otherwise} \end{cases}$$

$$AS_3 = \begin{cases} 0.5 + \frac{max(s_i, maxProb)}{2}, & \text{if correct} \\ 0.5 - \frac{min(s_i, maxProb)}{2}, & \text{otherwise} \end{cases}$$

Training Calibrator as Confidence Estimator:
- We extract features, namely, lengths, Semantic Textual Similarity (STS) value, number of common words between given sentences, and presence of negation words / numbers from the held-out instances to train the calibrator model.
- These features along with maxProb and prediction outputted by the model serve as inputs for the calibrator. Finally, we use a simple random forest implementation of Scikit-learn to train our calibrator that learns strong representations for the inputs.

Experiments and Results

Method	SNLI	MNLI			Stress Test			
		Matched	Mismatched	Avg	Competence	Distraction	Noise	Avg
MaxProb (AUC)	2.78	14.00	14.44	14.22	47.87	26.49	20.34	31.57
Calib T (%)	-181.2	-129.55	-127.86	-128.69	-48.65	-81.3	-91.17	-68.93
Calib C (%)	+8.97	+2.15	-1.36	+0.40	-3.75	**+8.27**	-0.80	+0.55
Proposed (%)	**+15.81**	**+2.35**	**+2.04**	**+2.19**	**+8.01**	+6.60	**+0.22**	**+5.64**

Table 1: Comparing percentage improvement of various calibration approaches on AUC of risk-coverage curve (over MaxProb) in in-domain (SNLI) and out-of-domain settings (MNLI, Stress Test) for the natural language inference (NLI) task.

Method	MRPC	QQP
MaxProb (AUC)	6.13	40.46
Calib T (%)	-148.87	+2.21
Calib C (%)	-0.82	+2.0
Proposed (%)	**+6.19**	**+13.9**

Table 2:Comparing % improvement of various calibration approaches on AUC of risk-coverage curve in IID (MRPC) and OOD (QQP) settings for Duplicate Detection task.

MaxProb Struggles in OOD setting:
- MaxProb performs well in the IID setting as it achieves low AUC values (2.78 on SNLI and 6.13 on MRPC). However, it fails to translate that in OOD (AUC of 14.22 on MNLI, 31.57 on Stress Test, and 40.46 on QQP).
- This implies that the model makes a significant number of incorrect predictions with relatively high MaxProb and thus needs to be calibrated.

Proposed Method Outperforms Other Methods:
- It achieves 15.81% and 6.19% improvement in the IID setting on SNLI and MRPC respectively. Furthermore, it achieves 2.19% on MNLI, 5.64% on Stress Test, and 13.9% on QQP in the OOD.
- Calib T degrades performance in both IID and OOD settings. However, Calib C results in a minor improvement in the IID setting (8.97% for SNLI) but does not consistently improve in the OOD setting (especially on MNLI Mismatched and Competence Test).