Appendix: Likelihood Ratios and Generative Classifiers for Unsupervised Out-of-Domain Detection In Task Oriented Dialog

Varun Gangal[1, 2 *], **Abhinav Arora**[2], **Arash Einolghozati**[2], **Sonal Gupta**[2]
[1] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Facebook Conversational AI, Menlo Park, CA 94303
vgangal@andrew.cmu.edu {abhinavarora,arashe,sonalg}@fb.com

Abstract

We present here the appendix section to our main paper Herein, we present some clarifications and derivations mentioned in the paper. We also present certain analysis and results which we had to exclude from the main paper owing to paucity of space.

1 Appendix

Maximizing $H(P)$ vs $-KL(P|\mathcal{U})$

$$\begin{aligned} KL(P|\mathcal{U}) &= \sum_{i=1}^{i=|L|} p_i \log \frac{p_i}{\mathcal{U}_i} \\ &= \sum_{i=1}^{i=|L|} p_i \log p_i - \sum_{i=1}^{i=|L|} p_i \log \mathcal{U}_i \\ &= -H(P) - \sum_{i=1}^{i=|L|} p_i \log \mathcal{U}_i \\ &= -H(P) - \log \frac{1}{|L|} \sum_{i=1}^{i=|L|} p_i \\ &= -H(P) - \log \frac{1}{|L|} \\ &= -H(P) + \log |L| \end{aligned}$$

Now,

$$\begin{aligned} \operatorname*{argmin}_i KL(P|\mathcal{U}) &= \operatorname*{argmin}_i (-H(P) + \log |L|) \\ \operatorname*{argmin}_i KL(P|\mathcal{U}) &= -\operatorname*{argmax}_i (H(P) - \log |L|) \\ \operatorname*{argmin}_i KL(P|\mathcal{U}) &= -\operatorname*{argmax}_i H(P) \end{aligned}$$

Hence, we can see that minimizing $-KL(P|U)$ is equivalent to maximizing $H(P)$.

[*] Work done by author while interning at Facebook Conv AI

Clarification about MSP

Our numbers for MSP baseline on SNIPS significantly differ from those reported in (Lin and Xu 2019), where MSP performs very poorly. The reason for this is that , differing slightly from the original method stated in (Hendrycks and Gimpel 2017), they did not tune the threshold on η on the validation set, and instead used a default threshold of 0.5 (We confirmed that this was the case through email correspondence with the authors). This naturally leads to much worse performance for MSP, since , as noted by (Hendrycks and Gimpel 2017) too, the maximum probability value in a softmax for *OOD* examples is only relatively lower compared to *ID* examples and is by itself still high [> 0.9] .

SNIPS,25% Results

Why Likelihood Ratio Works - A Longer Derivation

$$LLR_{\mathcal{M},\mathcal{B}}(X) = \frac{\hat{P}_M(X)}{\hat{P}_B(X)}$$

$$LLR_{\mathcal{M},\mathcal{B}}(X) = \frac{\Pi_{i=1}^{i=|S|} \hat{P}_{\mathcal{M}}(x_i|{x_1}^i)}{\Pi_{i=1}^{i=|S|} \hat{P}_{\mathcal{B}}(x_i|{x_1}^i)}$$

$$\log LLR_{\mathcal{M},\mathcal{B}}(X) = \sum_{i=1}^{i=|S|} \log \hat{P}_{\mathcal{M}}(x_i|x_1^i) - \log \hat{P}_{\mathcal{B}}(x_i|x_1^i)$$

We now continue the derivation here largely reproducing the logic from (Ren et al. 2019). Assume some word types are "background" i.e, they persist in the training data even after perturbation. Let us denote this set of background word types by W_B. The probabilities for these word types will be the same whether one uses $\mathcal{M}$ or $\mathcal{B}$. Hence, if $x_i \in W_B$, $\log \hat{P}_{\mathcal{M}}(x_i|x_1^i) = \log \hat{P}_{\mathcal{B}}(x_i|x_1^i)$

$$\log LLR_{\mathcal{M},\mathcal{B}}(X) = \sum_{i=1, x_i \notin W_B}^{i=|S|} \log \frac{\hat{P}_{\mathcal{M}}(x_i|x_1^i)}{\hat{P}_{\mathcal{B}}(x_i|x_1^i)}$$

As we can see, the LLR function is independent of background word types, and is only influenced by non-background or "semantic" word types. This derivation is obviously a bit of a simplification, since "background" or

Dataset	Model	$F_1 \uparrow$	$FPR@95\%TPR \downarrow$	AUROC $\uparrow$	$AUPR_{OOD} \uparrow$
SNIPS,25%	MSP,$\tau = 1e^3$	91.23 ± 3.07	40.10 ± 20.52	90.04 ± 5.49	94.28 ± 4.02
	$-KL(P\|\mathcal{R}) \approx -KL(P\|\mathcal{U})$	91.54 ± 3.13	36.90 ± 18.36	90.71 ± 5.16	94.71 ± 3.51
	LOF+LMCL	90.21 ± 3.84	42.61 ± 25.81	89.75 ± 7.16	95.15 ± 2.94
	$\mathcal{L}_{gen}$	91.13 ± 3.04	37.50 ± 15.44	90.42 ± 4.12	95.49 ± 1.88
	$\mathcal{L}_{gen}$+BACKLM+UNIFORM	94.87 ± 1.56	15.30 ± 7.04	96.34 ± 1.34	98.36 ± 0.52
	$\mathcal{L}_{gen}$+BACKLM+UNIGRAM	$\mathbf{96.28 \pm 1.75}$	$\mathbf{9.30 \pm 6.96}$	$\mathbf{98.12 \pm 1.03}$	$\mathbf{99.20 \pm 0.42}$
	$\mathcal{L}_{gen}$+BACKLM+UNIROOT	95.62 ± 1.66	13.40 ± 7.04	97.33 ± 1.08	98.84 ± 0.42

Table 1: Performance of the baseline methods and our proposed models on SNIPS,25%. $\downarrow$ ($\uparrow$) indicates lower (higher) is better. We can see that the $\mathcal{L}_{gen}$+BACKLM+<*Noise*> (where <*Noise*> is one of three noising schemes) approaches outdo their non LLR counterparts on most measures. For SNIPS, $-KL(P|\mathcal{R}) \approx -KL(P|\mathcal{U})$ since the training set is almost evenly distributed between the *ID* classes

"non-background" behaviour may not be tokenwise and may take place at the level of higher n-grams.

More Qualitative Examples

We present additional qualitative examples for each of the qualitative categories we identified in our main paper.

Category	Example	% in data
Overtly Powerful Action	1. order Mexican takeout 2. Send Mom $20 for the gas money 3. what is my checking balance 4. toast my bread 5. are my neighbors home yet?	20.55%
Action Memory	1. When was my last deposit from work? 2. are we out of bottled water 3. did I buy eggs last week 4. what is the balance of my loan 5. when was my favorite app last updated?	12.24%
Declarative Statement	1. I fix my own television when it's need to be fix 2. Cheer me on! 3. can you keep a secret 4. Am I your friend 5. make this for leadership	8.74%
Underspecified Query	1. what third chef job are available 2. Is the train running normally today? 3. find purple feather in Pinterest 4. show me the add on emperps via website 5. Ask dot com	33.94%
Speculative Question	1. Is Parks and Recreation having another season? 2. what is the release day for Xbox Scorpio? 3. which movies are being release on DVD this month 4. when does the next season of Aerial start 5. when is Empire on	6.91%
Subjective Question	1. What is the quickest way to learn Japanese? 2. what color glasses should I wear on a hot summer day? 3. how to redo your bathroom 4. how do I save energy? 5. How do I look?	27.99%

Table 2: We manually classify each *OOD* sentence in ROSTD into [1 or more] of 6 qualitative categories named self-explanatorily. Some additional examples in each of the categories w.r.t the similar table in the main paper.

A note on *ID* Classification Goodness

Note that our aim here is primarily to build a model which is a good *OOD* vs *ID* detector, in particular one with low FPR at high TPR values for the *OOD* class. We do not necessarily require the same model to be a good classifier between the ID classes, although many of our approaches do use that supervision which is available at training time - once we have a good *ID* vs *OOD* model, it can always be used in tandem before another model specifically optimized to be good at distinguishing between the *ID* classes.

References

Hendrycks, D., and Gimpel, K. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Lin, T.-E., and Xu, H. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5491–5496. Florence, Italy: Association for Computational Linguistics.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; DePristo, M. A.; Dillon, J. V.; and Lakshminarayanan, B. 2019. Likelihood ratios for Out-of-Distribution Detection. *arXiv preprint arXiv:1906.02845*.