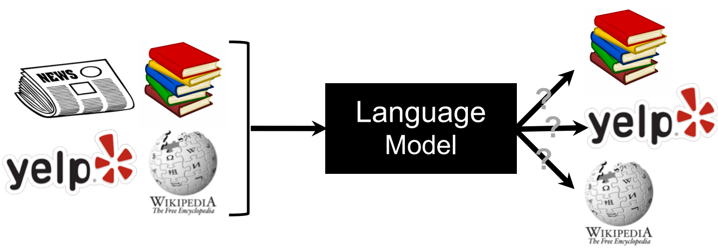
Distributionally Robust Language Modeling

Yonatan Oren*, Shiori Sagawa*, Tatsunori B. Hashimoto*, Percy Liang

Goal

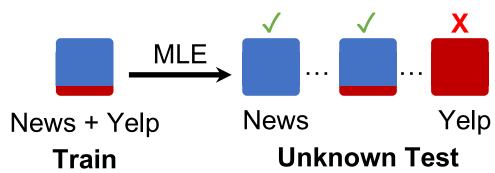
Given multi-domain training text, learn a language model that performs well on unknown test distributions



Problem

With standard training, a model performs *worse with more data* from outside the target domain.

	Yelp perplexity
Train on Yelp	32
Train on Yelp + News	45



Tool: Distributionally Robust Optimization

To achieve low loss on an unknown test distribution, we optimize the loss on the worst-case test distribution.

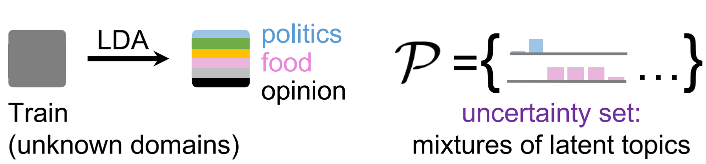
$$\underbrace{\mathbb{E}_{p_{\mathrm{x}}^{\text{test}}}[\ell(x;\theta)]}_{\text{unknown test loss}} \leq \underbrace{\sup_{p_{\mathrm{x}} \in \mathcal{P}} \mathbb{E}_{p_{\mathrm{x}}}[\ell(x;\theta)]}_{\text{worst-case test loss}}$$

if $p_{\mathrm{x}}^{\text{test}}$ = (unknown test distribution) is in $\mathcal{P} = \{ \ldots \}$ (uncertainty set: set of potential distributions)

Idea 1: Topic-Based Uncertainty Sets

Problem: Naïve, sentence-based uncertainty sets are too conservative. No domain information.

Solution: Define the uncertainty set by latent topics.



Result: Topics improve Yelp perplexity

	DRO no topics	DRO with topics
Train on Yelp + News	184	44

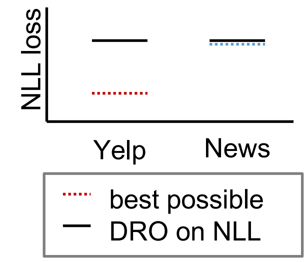
Idea 2: Baselined Loss

Problem: DRO on NLL loss overemphasizes hard topics

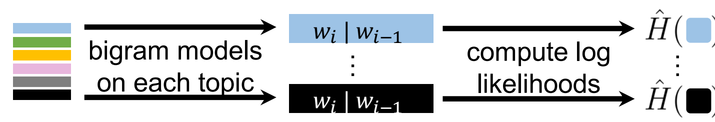
Solution: Define a loss baselined by topic difficulty

$$\ell((x,z);\theta) = -\log p_\theta(x) + \underbrace{H(z)}_{\text{Entropy; topic difficulty}}$$

- ✓ Accounts for topic difficulty
- ✓ Optimizes the distribution fit between the worst-case topic z and model, $\mathrm{KL}(p_{\mathrm{x}|z} \parallel p_\theta)$.



We estimate topic difficulty:



Result: Baselining improves Yelp perplexity

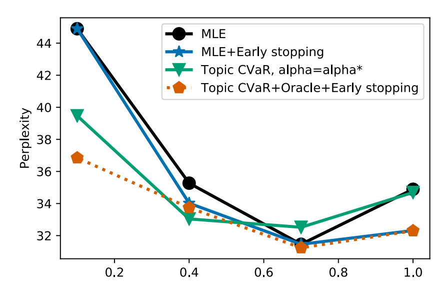
	DRO no baseline	DRO baselined
Train on Yelp + News	44	40

Experiments

Set-up: Train a Transformer on mixture of Yelp (α^*) and One Billion Words ($1-\alpha^*$).

	Yelp perplexity
MLE on	45
Topic DRO on	40

Minority training domain (Yelp): Topic DRO improves Yelp perplexity, estimating the *optimal* Yelp vs. news trade-off. Pure Yelp results (32) are impossible for our setting due to unknown test domain.



$P_{\mathrm{CVaR}} > P_{\mathrm{MLE}}$	$P_{\mathrm{MLE}} > P_{\mathrm{CVaR}}$
Huge servings, so plenty for leftovers. Every single person we spoke to on staff was absolutely incredible.	My girlfriend had an awful accident that hurt her leg & ankle which resulted in a fire and rescue ride. The address [PERSON] has listed is their old address.

Unseen domain (Trip Advisor): Topic DRO improves perplexity on Trip Advisor by 4 perplexity points, compared to standard training ($\alpha^* = 0.1$).

References

DRO: Ben-Tal+ 2013, Rockafelalr and Uryasev 2000, Duchi and Namkoong 2018
Topics in DRO: Hu+ 2018
Domain Adaptation: Shimodaira 2000, Pryzant+ 2017, Hoffman+ 2012