Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval

Devang Kulshreshtha[1,2], Robert Belfer[2], Iulian V. Serban[2], Siva Reddy[1,3]
Mila/McGill University[1] Korbit.AI[2] Facebook CIFAR AI Chair[3]

Paper Code + Data

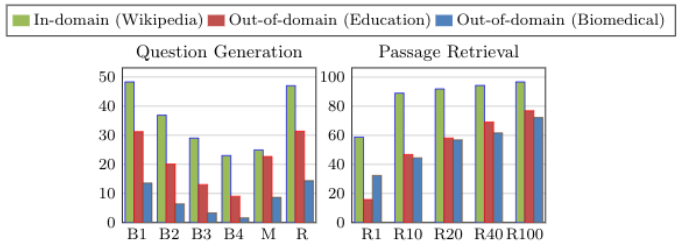
1 Supervised NLP models can fail drastically in *out-of-distribution* domains

2 We propose *Back-training* for Unsupervised Domain Adaptation (UDA) contrasting it with popular *Self-training* algorithm

3 Back-training *outperforms* Self-training on Question Generation (QG) and Passage Retrieval (IR) by a huge margin

1. Testing Out-of-Distribution Robustness

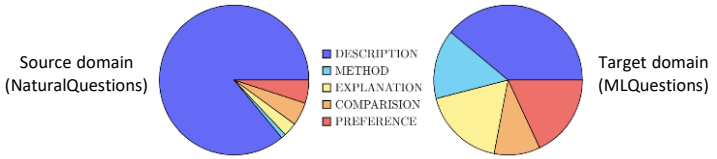
- Models trained on **Wikipedia** for Question Generation and Retrieval generalize poorly to domains like **Education** and **Bio-medical**



- Expensive to collect supervised data for each target domain
  - However, unsupervised data is cheap !
- **Goal:** Unsupervised Domain adaptation (UDA) by leveraging supervised source domain data and unsupervised target domain data
- **Task:** Question Generation (QG) and Passage Retrieval (IR)

MLQuestions: A New Benchmark Dataset

- Educational dataset consisting of Machine learning questions & articles for research in domain adaptation in QG and IR:
  - 35K unsupervised questions from Google search queries
  - 50K unsupervised passages from Wikipedia ML pages
  - 3K aligned question-passage pairs for model evaluation



- MLQuestions has higher diversity of questions, making UDA challenging

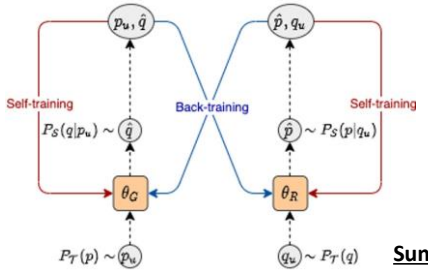
2. Unsupervised Domain Adaptation : Back-training vs Self-Training

Self-training : Question Generation

- First train source domain QG model $P_s(q|p)$
- Sample passages from target domain $p_u = P_T(passages)$
- Generate synthetic questions for target passages $q' = P_s(q|p_u)$
- Finetune on target inputs and synthetic outputs $P_T(q|p)$

Back-training : Question Generation

- Train source domain QG model $P_s(q|p)$ and retriever $P_s(p|q)$
- Sample questions from target domain $q_u = Q_T(questions)$
- Use retriever to find passage which can generate target question $p' = P_s(p|q_u)$
- Finetune on synthetic inputs and target outputs $P_T(p|q)$



Summary

☺ Real inputs sampled from target domain

☹ Synthetic outputs generated by source domain model

Summary

☹ Synthetic inputs generated by source domain model

☺ Real outputs sampled from target domain

HYPOTHESIS

Real outputs belonging to target domain should be more desirable than inputs having same properties for adaptation to target domain

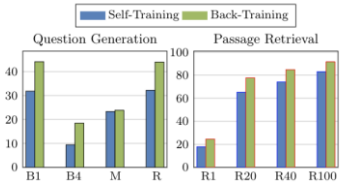
3. Results

Experimental Setup

- BART encoder-decoder to train Question Generation models
- Dense Passage Retriever based on BERT
- Source Domain (supervised data) : NaturalQuestions
- Target Domain (unsupervised data) : MLQuestions

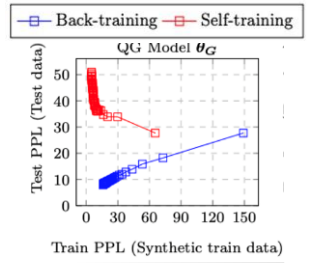
Results

- Back-training excels Self-training
  - 12 BLEU points on QG
  - 9 points on IR



Training Curves for Question Generation

- Self-training leads to overfitting
- Back-training generates data closer to target domain
- Back-training scales with amount of unlabeled data



Qualitative Results for Question Generation

Input Passage :

If the line is a good fit for the data then the residual plot will be random. However, if the line is a bad fit for the data then the plot of residuals will be random.

Output Questions :

No-adaptation: *What is the meaning of random plot in statistics?*
ST: *What is the meaning of random plot in statistics?*
BT: *How do you know if a residual plot is random?*
Reference: *How do you know if a residual plot is good?*