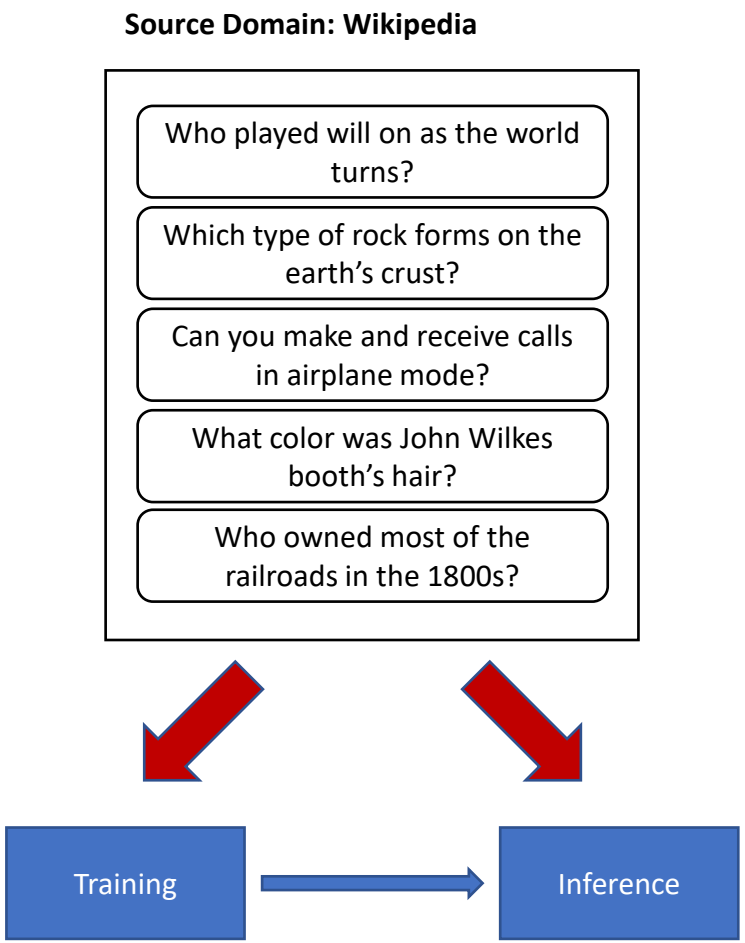
Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval

Devang Kulshreshtha[1,2] , **Robert Belfer**[2] , **Iulian Vlad Serban**[2] , **Siva Reddy**[1,3]

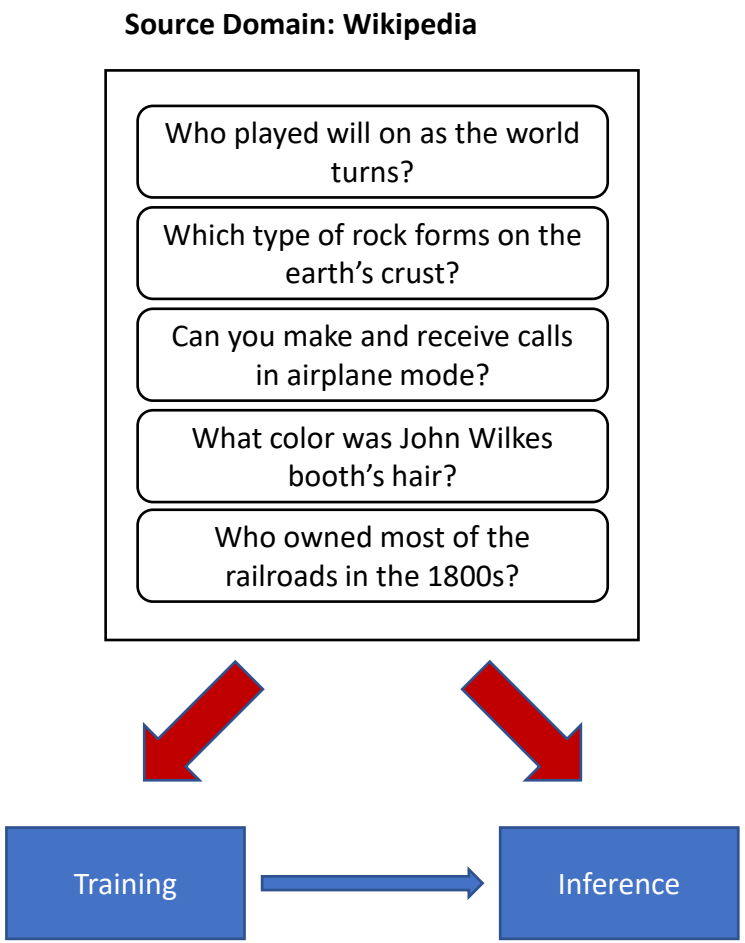
[1]Mila/McGill University [2]Korbit.AI [3]Facebook CIFAR AI Chair

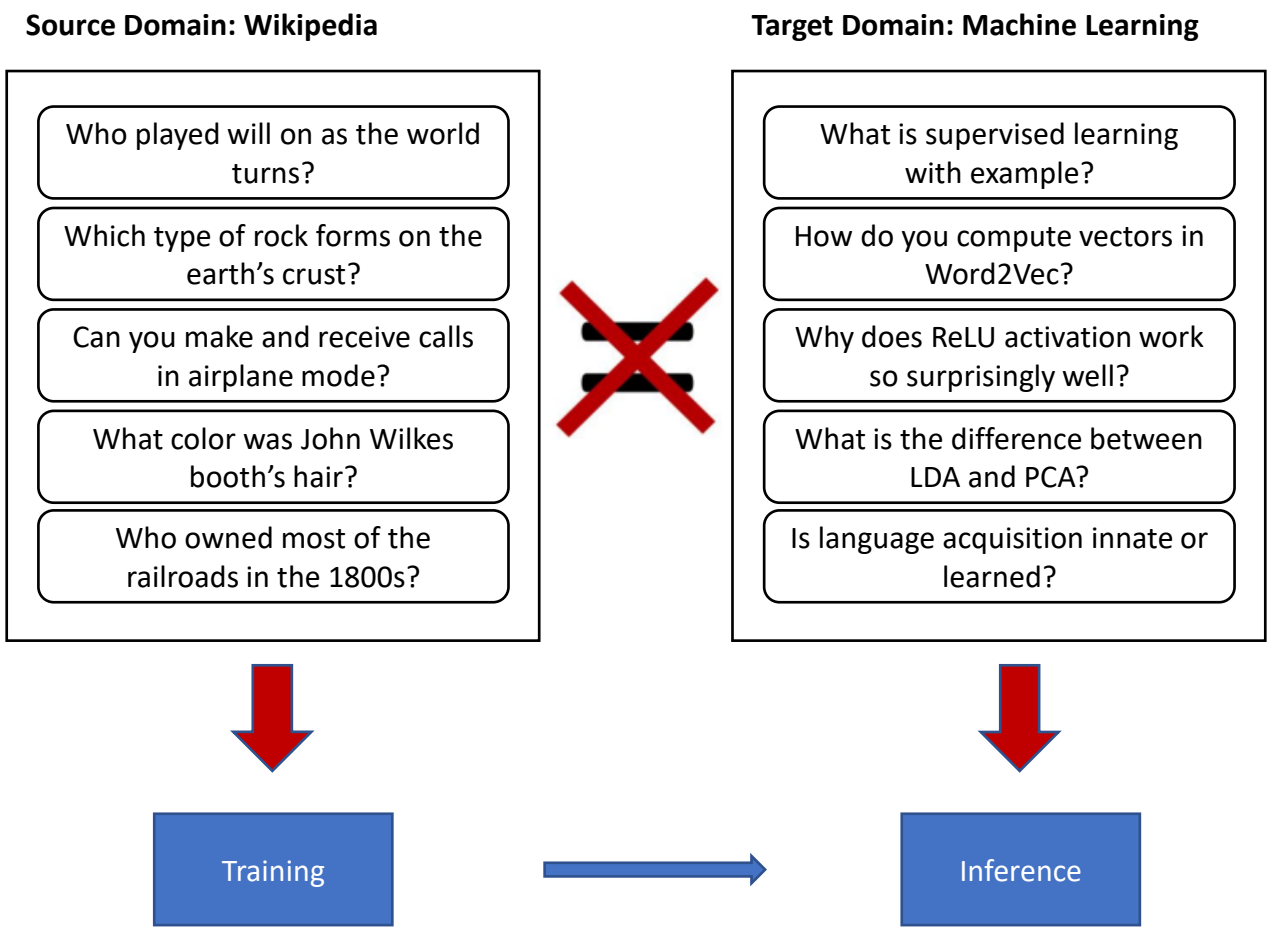
EMNLP 2021

Out-of-Distribution Robustness

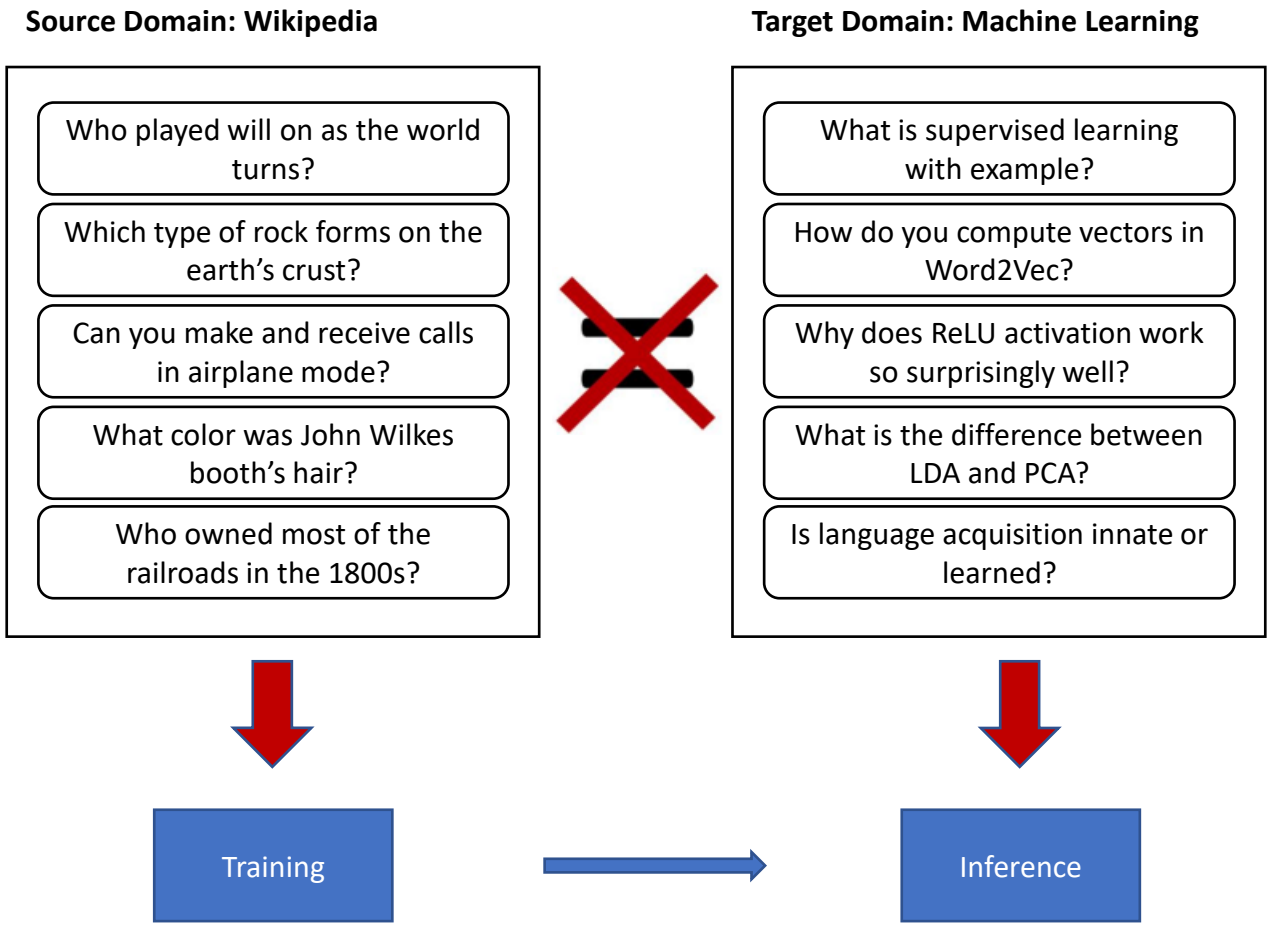


Out-of-Distribution Robustness

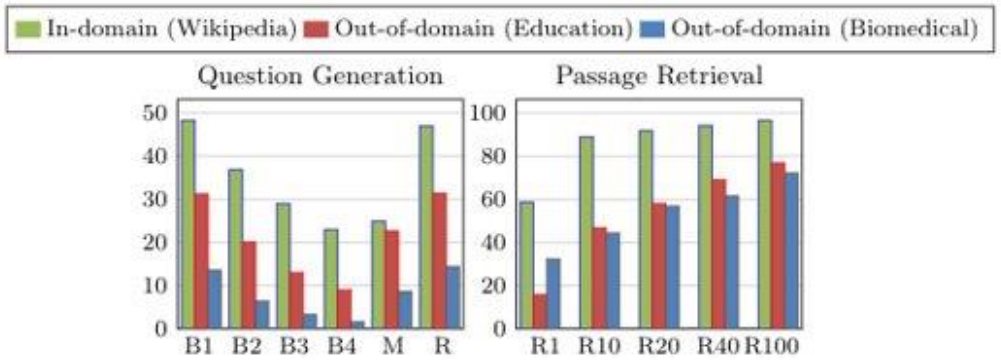




Out-of-Distribution Robustness



Out-of-Distribution Robustness



Source Domain: Wikipedia

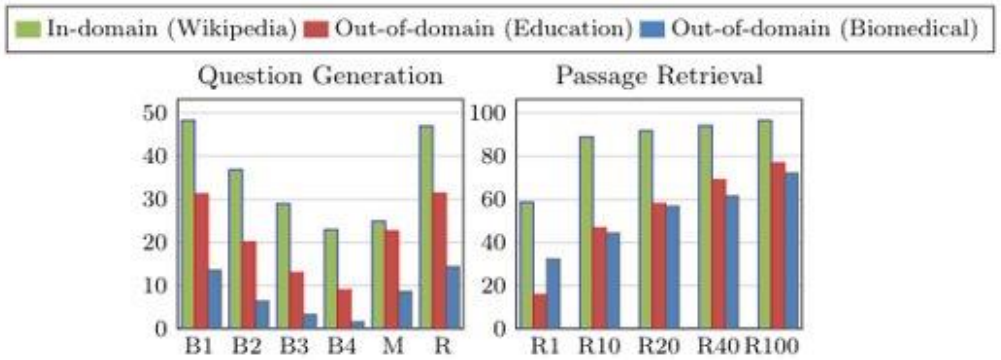
- Who played will on as the world turns?
- Which type of rock forms on the earth's crust?
- Can you make and receive calls in airplane mode?
- What color was John Wilkes booth's hair?
- Who owned most of the railroads in the 1800s?

Target Domain: Machine Learning

- What is supervised learning with example?
- How do you compute vectors in Word2Vec?
- Why does ReLU activation work so surprisingly well?
- What is the difference between LDA and PCA?
- Is language acquisition innate or learned?

Training → Inference

Out-of-Distribution Robustness



"Ok, I'll just collect supervised data for each domain I encounter!"

Source Domain: Wikipedia

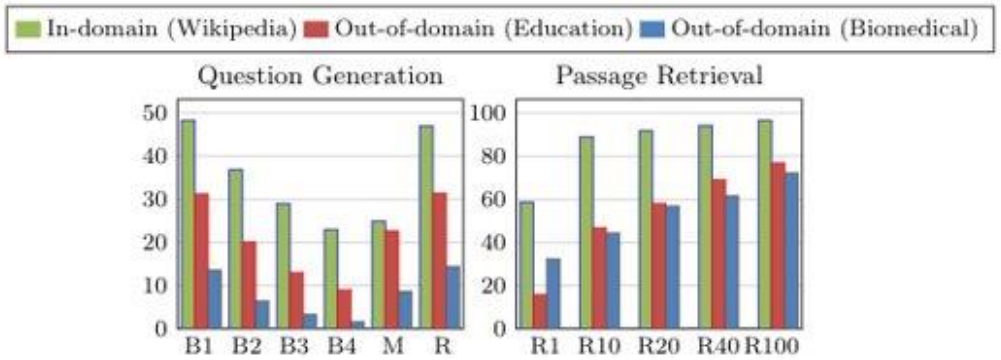
- Who played will on as the world turns?
- Which type of rock forms on the earth's crust?
- Can you make and receive calls in airplane mode?
- What color was John Wilkes booth's hair?
- Who owned most of the railroads in the 1800s?

Target Domain: Machine Learning

- What is supervised learning with example?
- How do you compute vectors in Word2Vec?
- Why does ReLU activation work so surprisingly well?
- What is the difference between LDA and PCA?
- Is language acquisition innate or learned?

Training → Inference

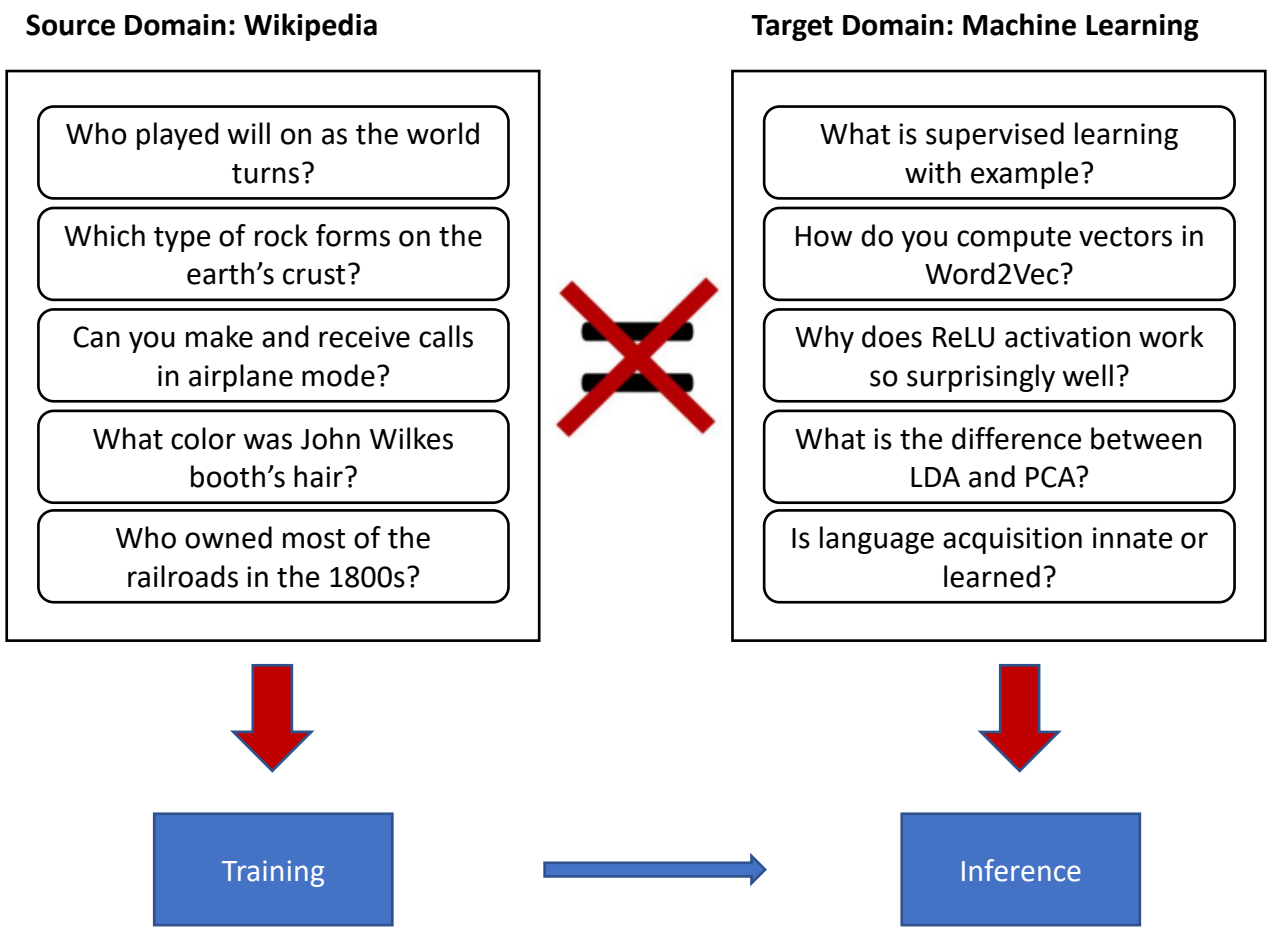
Out-of-Distribution Robustness



"Ok, I'll just collect supervised data for each domain I encounter!"

However, unsupervised domain data is cheap!

- Millions of passages from text books
- Millions of questions from real students on the internet (e.g. StackExchange)



Unsupervised Domain Adaptation (UDA)

- **GOAL:** Unsupervised Domain adaptation (UDA) by leveraging supervised source domain data and unsupervised target domain data

- **TASK:** Question Generation (QG) and Passage Retrieval (IR)

- **INPUT:**

  - Source domain aligned question-passage pairs $D_S = \{(q_s^i, p_s^i)\}$
  - Target domain unaligned questions $Q_u = \{(p_s^i)\}$
  - Target domain unaligned passages $P_u = \{(p_s^i)\}$

- **OUTPUT:**

  - Target domain QG model $P_T(q|p)$
  - Target domain IR model $P_T(p|q)$

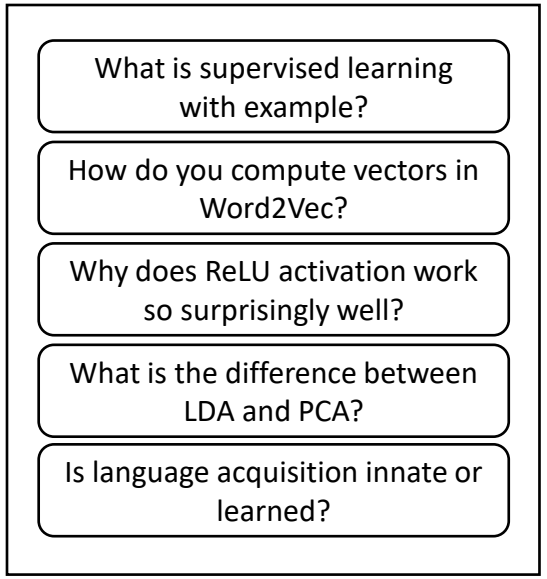
MLQuestions: A New Benchmark Dataset

- 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: A New Benchmark Dataset

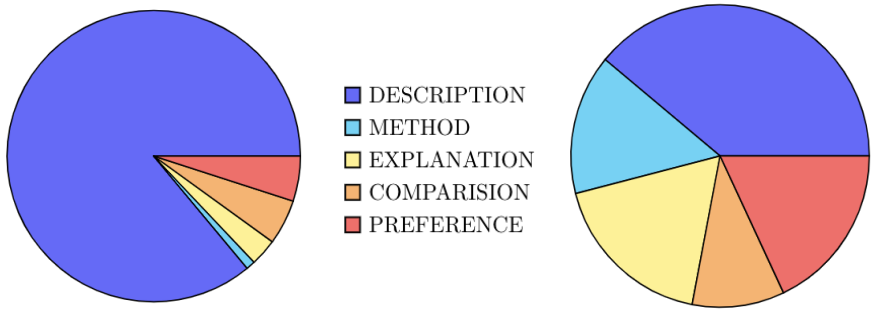
- Educational dataset consisting of Machine learning questions & articles for research in domain adaptation in QG and IR:
  - 35K unsupervised questions from Google search queries
  - 50K unsupervised passages from Wikipedia ML pages
  - 3K aligned question-passage pairs for model evaluation

Target Domain: MLQuestions



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



NaturalQuestions (source domain) MLQuestions (target domain)

Target Domain: MLQuestions

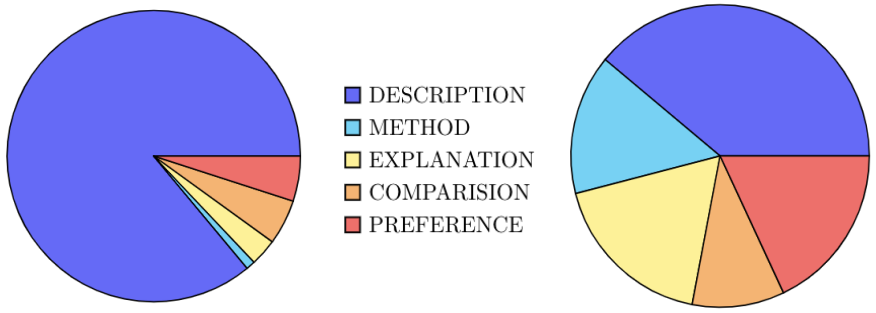
- What is supervised learning with example?
- How do you compute vectors in Word2Vec?
- Why does ReLU activation work so surprisingly well?
- What is the difference between LDA and PCA?
- Is language acquisition innate or learned?

Source Domain: NaturalQuestions

- Who played will on as the world turns?
- Which type of rock forms on the earth's crust?
- Can you make and receive calls in airplane mode?
- What color was John Wilkes booth's hair?
- Who owned most of the railroads in the 1800s?

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



NaturalQuestions (source domain) MLQuestions (target domain)

Target Domain: MLQuestions

- What is supervised learning with example?
- How do you compute vectors in Word2Vec?
- Why does ReLU activation work so surprisingly well?
- What is the difference between LDA and PCA?
- Is language acquisition innate or learned?

Source Domain: NaturalQuestions

- Who played will on as the world turns?
- Which type of rock forms on the earth's crust?
- Can you make and receive calls in airplane mode?
- What color was John Wilkes booth's hair?
- Who owned most of the railroads in the 1800s?

- MLQuestions has higher diversity of questions, making UDA challenging !

Unsupervised Domain Adaptation Algorithm

Unsupervised Domain Adaptation Algorithm

Self-training : Question Generation

Unsupervised Domain Adaptation Algorithm

Self-training : Question Generation

- First train source domain QG model $P_s(q|p)$

Unsupervised Domain Adaptation Algorithm

Self-training : Question Generation

- First train source domain QG model $P_s(q|p)$
- Sample passages from target domain $p_u = P_T(passages)$

Unsupervised Domain Adaptation Algorithm

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 $q' = P_s(q|p_u)$

Unsupervised Domain Adaptation Algorithm

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 $P_T(q|p)$

Unsupervised Domain Adaptation Algorithm

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 $P_T(q|p)$

Summary

☺ Real inputs sampled from target domain

☹ Synthetic outputs generated by source domain model

Unsupervised Domain Adaptation Algorithm

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

Summary

☺ Real inputs sampled from target domain

☹ Synthetic outputs generated by source domain model

Unsupervised Domain Adaptation Algorithm

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 model $P_s(p|q)$

Summary

☺ Real inputs sampled from target domain

☹ Synthetic outputs generated by source domain model

Unsupervised Domain Adaptation Algorithm

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 model $P_s(p|q)$
- Sample questions from target domain $q_u = Q_T(questions)$

Summary

☺ Real inputs sampled from target domain

☹ Synthetic outputs generated by source domain model

Unsupervised Domain Adaptation Algorithm

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 model $P_s(p|q)$
- Sample questions from target domain $q_u = Q_T(questions)$
- Use retriever model to find passage which can generate target question $p' = P_s(p|q_u)$

Summary

☺ Real inputs sampled from target domain

☹ Synthetic outputs generated by source domain model

Unsupervised Domain Adaptation Algorithm

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 model $P_s(p|q)$
- Sample questions from target domain $q_u = Q_T(questions)$
- Use retriever model 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

Unsupervised Domain Adaptation Algorithm

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 $P_T(q|p)$

Summary

☺ Real inputs sampled from target domain

☹ Synthetic outputs generated by source domain model

Back-training : Question Generation

- Train source domain QG model $P_s(q|p)$ and retriever model $P_s(p|q)$
- Sample questions from target domain $q_u = Q_T(questions)$
- Use retriever model 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

☹ Synthetic inputs generated by source domain model

☺ Real outputs sampled from target domain

Unsupervised Domain Adaptation Algorithm

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 $P_T(q|p)$

Summary

☺ Real inputs sampled from target domain

☹ Synthetic outputs generated by source domain model

Back-training : Question Generation

- Train source domain QG model $P_s(q|p)$ and retriever model $P_s(p|q)$
- Sample questions from target domain $q_u = Q_T(questions)$
- Use retriever model 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

☹ 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

Experimental Setup

- BART encoder-decoder to train Question Generation model
- Dense Passage Retriever based on BERT
- Source Domain - NaturalQuestions (Wikipedia domain)
- Target Domain
    I. MLQuestions (Education Domain)
    II. PubMedQA (Biomedical Domain)

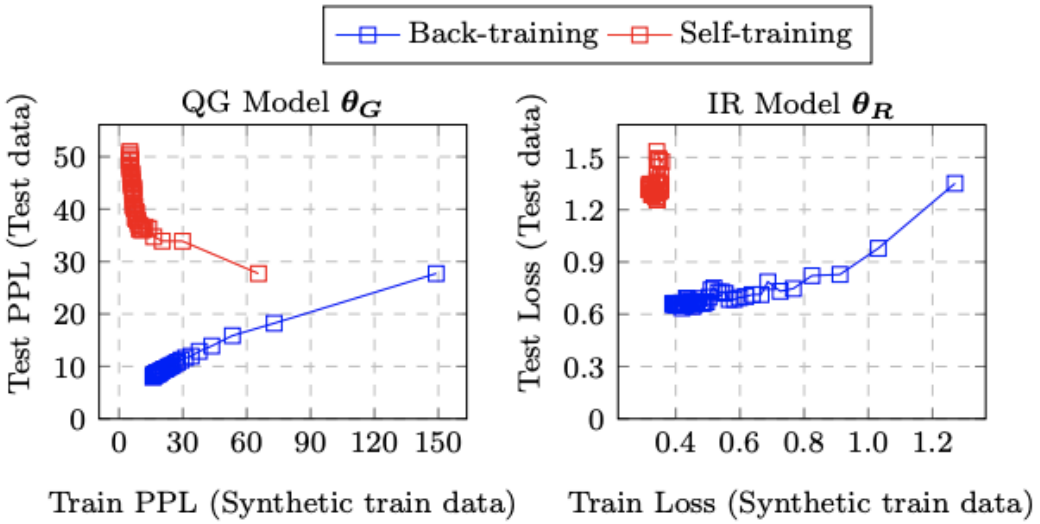
Results

- Back-training outperforms Self-training across **education** and **biomedical** domains
  - 12 BLEU points on Question Generation
  - 9 points on Passage Retrieval

Dataset	Model	Question Generation						Passage Retrieval			
		B1	B2	B3	B4	M	R	R@1	R@20	R@40	R@100
MLQuestions	No-adaptation	31.23	20.07	13.05	9.04	22.70	31.38	15.86	58.13	69.13	76.86
	Self-Training	31.81	20.74	13.61	9.43	23.31	32.18	17.86	65.26	74.13	83.06
	Back-Training	**44.12**	**32.86**	**24.21**	**18.48**	**23.83**	**43.97**	**24.53**	**77.73**	**84.8**	**91.73**
PubMedQA	No-adaptation	13.57	6.41	3.31	1.62	8.67	14.38	32.4	56.8	61.6	72.2
	Self-Training	13.36	6.28	3.25	1.64	8.84	15.00	32.8	57.0	63.6	72.8
	Back-Training	**26.71**	**17.01**	**11.80**	**8.25**	**16.99**	**25.14**	**55.4**	**79.8**	**81.8**	**85.8**

Table 4: Results of unsupervised domain adaptation. *No-adaptation* denotes the model trained on NaturalQuestions and tested directly on MLQuestions/PubMedQA without any domain adaptation.

Training Curves



Observations:

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

Qualitative results for Question Generation

MLQuestions

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 :

**ST = Self-Training* **BT = Back-Training*

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?*

Conclusions

- Supervised NLP models can fail under OOD generalization
- In self-training, inputs are from target domain (real) and outputs are noisy (predicted)
- In back-training, inputs are noisy (predicted) and outputs are from target domain (real)
- Self-training can increase overfitting to source domain
- Back-training generates data closer to target domain

Paper

Code + Data

McGill UNIVERSITY

KORBIT

Mila