EMNLP-IJCNLP 2019

Adversarial Domain Adaptation for Machine Reading Comprehension

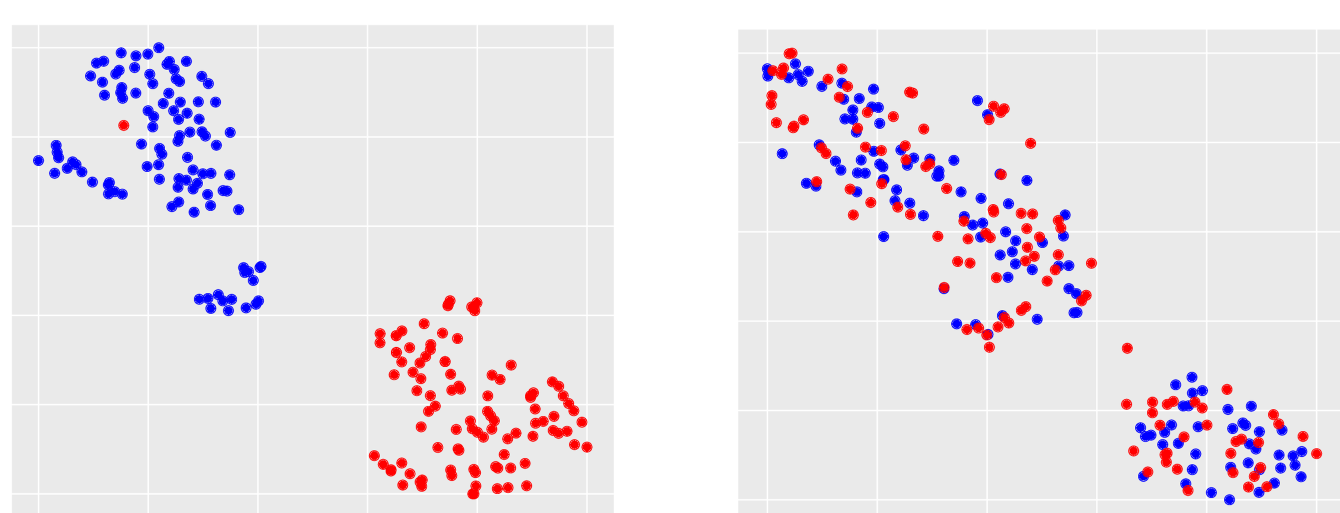
Huazheng Wang[1], Zhe Gan[2], Xiaodong Liu[3], Jingjing Liu[2], Jianfeng Gao[3], Hongning Wang[1]

[1]University of Virginia [2]Microsoft Dynamics 365 AI Research [3]Microsoft Research

Motivation & Contribution

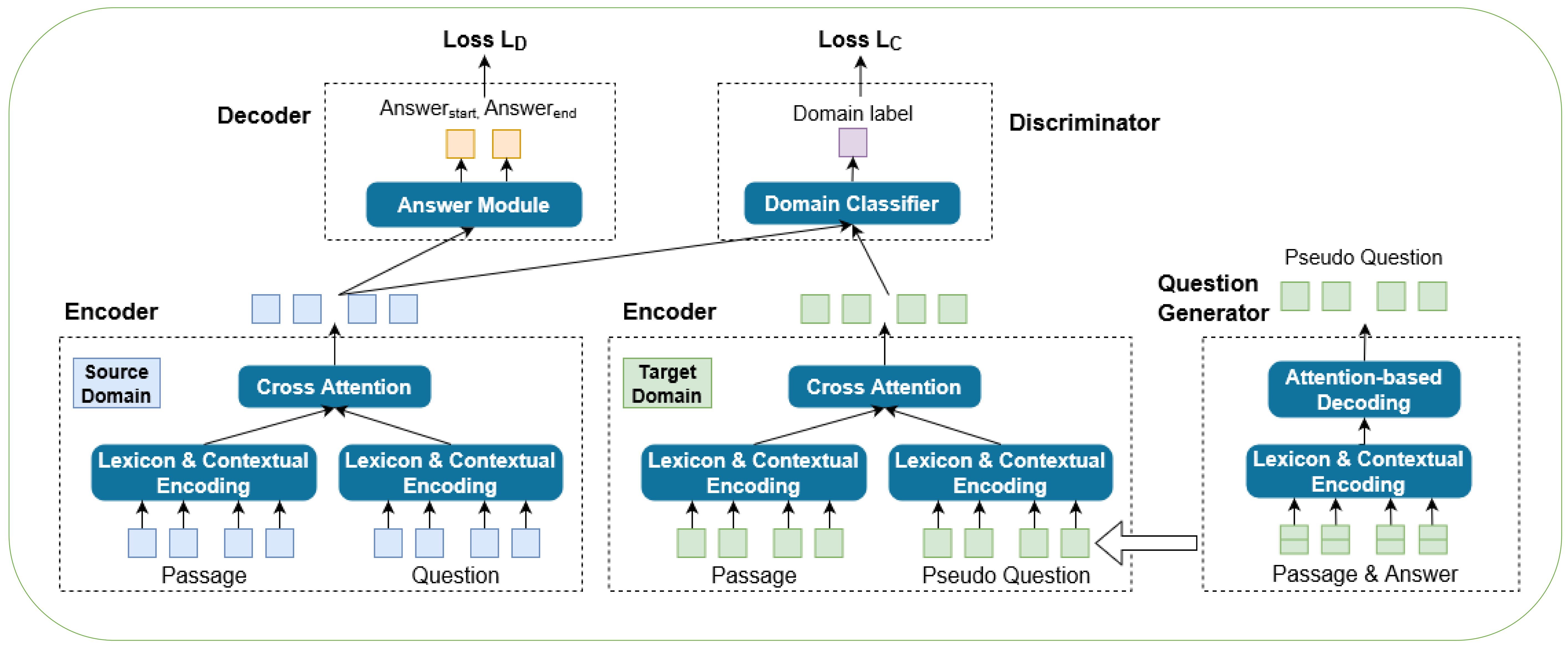
- Motivation
  - Recent success in MRC relies on large-scale annotated in-domain data (e.g., SQuAD)
  - Directly adapting models from source domain to low-resource target domain performs poorly due to domain shift
- Contribution
  - Unsupervised Domain Adaptation by generating pseudo data on target domain and learning *domain-invariant* representations through adversarial learning

T-SNE plot of encoded feature representations



Without domain adaptation — With domain adaptation

AdaMRC Framework



- Question Generator (QG): using passage and answer (extracted by NER) as input for generating pseudo questions in the target domain
- Encoder & Decoder: source domain and target domain share the same encoder & decoder (i.e., MRC Module)
- Discriminator: an MLP as domain classifier. A gradient reverse layer is used for gradient backpropagation
- Training: pseudo target domain data is used for encoder and discriminator training, but ***not*** for decoder training, because the synthetic data could be noisy

Training Algorithm

Algorithm 1 AdaMRC training procedure.

Input: source domain labeled data $S = \{p^s, q^s, a^s\}$, target domain unlabeled data $T = \{p^t\}$

Train the MRC model $\theta^s = (\theta_e^s, \theta_d^s)$ on source domain S;

Train the QG model θ_{QG} on source domain S;

Generate $T_{gen} = \{p^t, q^t, a^t\}$ using the QG model;

Initialize $\theta = (\theta_e, \theta_d, \theta_c)$ with θ^s;

for epoch $\leftarrow 1$ to #epochs **do**

  Optimize θ on $S \cup T_{gen}$. Each minibatch is composed with k_s samples from S and k_t samples from T_{gen};

end for

Output: Model with the best performance on the target development set θ^*.

Experimental Results

Dataset	Domain
SQuAD (v1.1)	Wiki
NewsQA	News
MS MARCO (v1)	Web

- Main results are based on Stochastic Answer Network (SAN)
- AdaMRC consistently improves performance over baselines
- Direct data augmentation and fine-tuning (SynNet) hurts performance
- Question generation is effective (margin with "AdaMRC with GT questions" is relatively small)
- Generalizable to other datasets and other MRC models with consistent performance gain

Method	EM/F1
SQuAD ⟶ NewsQA	
SAN	36.68/52.79
SynNet + SAN	35.19/49.61
AdaMRC	38.46/54.20
AdaMRC with GT questions	39.37/54.63
NewsQA ⟶ SQuAD	
SAN	56.83/68.62
SynNet + SAN	50.34/62.42
AdaMRC	58.20/69.75
AdaMRC with GT questions	58.82/70.14
SQuAD ⟶ MS MARCO (BLEU-1/ROUGE-L)	
SAN	13.06/25.80
SynNet + SAN	12.52/25.47
AdaMRC	14.09/26.09
AdaMRC with GT questions	15.59/26.40
MS MARCO ⟶ SQuAD	
SAN	27.06/40.07
SynNet + SAN	23.67/36.79
AdaMRC	27.92/40.69
AdaMRC with GT questions	27.79/41.47

- Compatible with pre-trained language models

Method	EM/F1
SAN	36.68/52.79
AdaMRC + SAN	38.46/54.20
SAN + ELMo	39.61/55.18
AdaMRC + SAN + ELMo	40.96/56.25
BERT$_{base}$	42.00/58.71
AdaMRC + BERT$_{base}$	42.59/59.25

- Can be extended to semi-supervised setting

Ratio (%Labeled data)	SAN	AdaMRC+ SAN
0%	36.68/52.79	38.46/54.20
5%	47.61/62.69	48.50/63.17
10%	48.66/63.32	49.64/63.94
20%	50.75/64.80	51.14/65.38
50%	53.24/67.30	53.34/67.30
100%	56.48/69.14	56.29/68.97