Robust Q&BAE: Improving Out-of-Domain Question Answering Performance With Data Augmentation Techniques Inspired by Adversarial Perturbation Methods

By Lynn Kong, Philip Weiss, and Adam Pahlavan

Problem:

- Classical ML assumes training and test data come from the same distribution
  - In cases where this assumption does not hold, models can exhibit a large drop in accuracy on out-of-domain performance
  - The sub-field of ML which studies the task of improving generalization of models to out-of-domain data is known as **domain adaptation**
- **Data Augmentation:** An approach that generates synthetic data samples that look like they come from the out-of-domain distribution, and then use these samples in training (Fig. 1)
- We apply a Data Augmentation-based based on BERT-based token replacement as an approach to the the Robust QA track domain adaptation problem

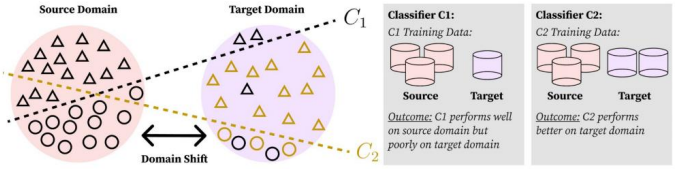

Figure 1. Data augmentation approach to domain adaptation for the training of a binary classifier. Adding augmented data samples improves the out-of-domain accuracy of the second classifier. Styles adapted from a figure in [1].

Background

- Several different categories of domain adaptation (Fig. 2):
  - **Unsupervised Methods:** Only have unlabeled data on out-of-domain dataset
  - **Supervised Methods:** Have access to labeled data on out-of-domain dataset
- The RobustQA problem is in the supervised category since we have access to a limited number of labeled out-of-domain samples at training time. This allowx us to use the (more powerful) supervised category of domain adaptation methods.
- Our specific approach to the problem is most closely similar to the **token perturbation** method that is in the **data-centric** and **rule-based** categories (Fig. 2)
- In particular, we choose to use a data augmentation approach that perturbs context paragraphs in our out-of-domain training set to generate new samples to train on
  - Our study investigates two types of perturbations: **BERT-based** and **synonym-based**
  - As a result, we can expand the size of our training dataset

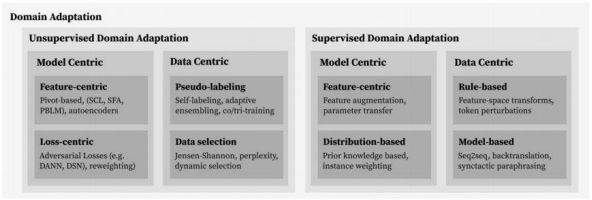

Figure 2. Overview of domain adaptation methods. The techniques studied here fall within the supervised domain adaptation / data-centric / rule-based category and are more closely related to token perturbation methods. This taxonomy was synthesized from the literature reviews in [2, 3, 4].

Methods & Experiments

High Level Approach

- Our approach to improving performance on out of domain datasets for the RobustQ&A default project is through dataset augmentation. Specifically, we are co-opting techniques from Siddhant et. al [5]] which uses BERT-based token replacements to generate adversarial examples.
- This class has a perturb method, which takes in as input a sentence S, and a tuple of answer starts (the label corresponding to the question answer in the training data). The goal of the perturb method is to output a "perturbed" version of the original sentence for which semantic similarity to the original sentence is preserved.
- Class **BERTDatasetAugmentation**- instantiated with a *language model* (synonyms or BERT MLM), *semantic similarity function (USE or SBERT)*, *k* (number of perturbations per sentence), and *number of indexes to consider.*

Training Procedure

- We tried ~14 combinations of languages models, semantic similarity functions, number of masks, and number of indexes to consider. Each model takes ~4 hours, and there are >200 hyperparam combinations. We ran our experiments in two phases- the first to pick the sim function and masked language model. Then, once those were fixed, we varied number of perturbations and number of indices. Overall, we trained **14 models (~50 hours)**
- **Phase 1:** Varied token unmasking method (Bert MLM vs. synonym scorer), as well as the similarity score function (Universal Sentence Encoder [USE] vs. Sentence Bert [SBERT]).
- **Phase 2:** Using phase 1 best- SBERT+MLM, k=1, 2, 3, 4, 5, while keepings perturbed samples fixed at 7% of data. Then, fixing k=3, and varying percentage of perturbed samples as 2.75, 5.15, 6.91, 9.81, 12.99% of training data (corresponds to number of indices to consider in our perturber class).

Results & Analysis

Model Name	Token Gen	Semantic Sim	F1	EM
Baseline-0.1	-	-	48.208	32.461
synonym+sbert*	Synonym	SBERT	47.927	33.77
synonym+use	Synonym	USE	50.424	36.649
bert-mlm+sbert	BERT-MLM	SBERT	**51.306**	**37.696**
bert-mlm+use	BERT-MLM	USE	49.43	36.39

Table 1. Validation performance of Phase 1 models.

Phase 1

- All non-baseline models performed on par or better than baseline (Table 1)
- Interestingly, Synonym performed better with USE, while BERT-MLM performed better with SBERT - likely due to shared BERT optimization space
- BERT-MLM+SBERT model performed best on F1 and EM
- BERT-MLM perturbed sentence made more semantic sense than Synonym (Fig. 4; Ex 1 and Ex 2)

(i) Ex. 1: original sentence

Question: Why did Ray Eberle die?
Context paragraph: Ray Eberle died of a heart attack in Douglasville, Georgia on August 25, 1979, aged 60.
Answer: heart attack

(ii) Ex. 1: Synonym perturbed sentence (1 mutation)

Question: Why did Ray Eberle die?
Context paragraph: Ray Eberle died of a heart blast in Douglasville, Georgia on August 25, 1979, aged 60.
Answer: heart blast

(iii) Ex. 1: BERT-MLM perturbed sentence (1 mutation)

Question: Why did Ray Eberle die?
Context paragraph: Ray Eberle died of a heart condition in Douglasville, Georgia on August 25, 1979, aged 60.
Answer: heart condition

(iv) Ex. 2: original sentence

Question: What business published NHL FaceOff 2003?
Context paragraph: NHL FaceOff 2003 is an ice hockey video game made by SolWorks and published by Sony Computer Entertainment of America, released on the PlayStation 2.
Answer: Sony Computer Entertainment

(v) Ex. 2: Synonym perturbed sentence (1 mutation)

Question: What business published NHL FaceOff 2003?
Context paragraph: NHL FaceOff 2003 is an ice hockey_game video game made by SolWorks and published by Sony Computer Entertainment of America, released on the PlayStation 2.
Answer: Sony Computer Entertainment

(vi) Ex. 2: Synonym perturbed sentence (1 mutation)

Question: What business published NHL FaceOff 2003?
Context paragraph: NHL FaceOff 2003 is an ice dance video game made by SolWorks and published by Sony Computer Entertainment of America, released on the PlayStation 2.
Answer: Sony Computer Entertainment

Figure 4. Examples of original, synonym-perturbed, and BERT-MLM- perturbed sentences for single mutations

Figure 3. Validation performance of Phase 2 models.

Phase 2

- Overall results were more nuanced
- 1 mutation performed better than multiple mutations (Fig. 3)
- Model performed best with ~7% of training data being perturbed samples (Fig. 3); index upper bound = 30)

Question: What is the name of the chromosome where you can find C14orf159?
Context Paragraph: UUPF0317 protein C14orf159, mitochondrial is a protein that in humans is encoded by the C14orf159 gene (**chromosome 14** open reading frame 159).
True Answer: chromosome 14
Baseline Model Answer: mitochondrial is a protein that in humans is encoded by the C14orf159 gene (**chromosome 14**
BERT-MLM+SBERT Model Answer: chromosome 14

Figure 4. Example where BERT+MLM+SBERT model properly classifies but baseline does not.

- Baseline misclassifies C14orf159 protein example but BERT+MLM+SBERT model properly classifies (Fig. 4)
- We hypothesize that since the `relation_extraction` training dataset contains several examples regarding genes and chromosomes,the reason for the BERT-MLM-SBERT model's success is the ability to perturb and expand these sentences into a wider variety to learn from when compared to the baseline

Limitations

(i) Ex. 3: original sentence

Question: What instrument is Flute sonata in E minor (HWV 379) for?
Context paragraph: The Flute sonata in E minor (HWV 379) was composed (circa 1727-28) by George Frideric Handel for flute and keyboard (harpsichord).
Answer: flute

(ii) Ex. 3: BERT-MLM perturbed sentence (3 mutations)

Question: What instrument is Flute sonata in E minor (HWV 379) for?
Context paragraph: The Flute sonata in E minor (HWV 379) was composed (circa 1727-28) by George Frideric Handel for keyboard. and harpsichord clue
Answer: keyboard

(iii) Ex. 4: original sentence

Question: Who was the brother of Eadwulf Cudel?
Context paragraph: Uhtred the Bold and Eadwulf Cudel were sons of Waldheof, earl of Bamburgh, who died in 1006.
Answer: Uhtred the Bold

(iv) Ex. 4: BERT-MLM perturbed sentence (3 mutations)

Question: Who was the brother of Eadwulf Cudel?
Context paragraph: Uhtred the Bold and oswald Cudel were brothers of Waldheof, king of Bamburgh, who died in 1006.
Answer: Uhtred the Bold

(v) Ex. 5: original sentence

Question: The publisher that published The Case of the Late Pig is what?
Context paragraph: The Case of the Late Pig is a crime novel by Margery Allingham, first published 1937, by Hodder & Stoughton.
Answer: Hodder & Stoughton

(vi) Ex. 5: BERT-MLM perturbed sentence (3 mutations)

Question: The publisher that published The Case of the Late Pig is what?
Context paragraph: The Case of the Late ##wife is a horror story by Margery Allingham, first published 1937, by Hodder & Stoughton.
Answer: Hodder & Stoughton

Figure 5. Example of augmentation with 3 mutations.

We chose to forgo certain implementations or axis of investigation in favor of running more complete experiments.

- No token ranking by importance, instead just limit mutation token candidates to nouns
- Inconsistent tokenization - split sentence into word level tokens, but BERT-MLM tokenizer is at subword level
- Limited mutations to the first 200 words due to BERT-MLM 512-input length limit
- Only mutated context paragraph, led to question and context inconsistency as seen in Ex 4 and Ex 5 in Fig. 5.

Conclusion

We implemented a data augmentation pipeline with multiple semantic and perturbation configuration parameters, and successfully demonstrated that augmented data from this pipeline increases model performance on low-resource Q&A. We observed that models trained on BERT-MLM and SBERT-scorer augmented datasets performed best, which deviated from the original BAE paper that used BERT-MLM and USE-scorer. We also investigated how the level of perturbation in the training set (number of mutations per sample and index upper bound), and found that 1 mutation and ~7% of training data being perturbed samples performed best. However, this latter result is less conclusive.

In the future, we want resolve the implementation limitations, such as creating a sliding window approach so token mutations can happen anywhere in the input text. Additionally, we want to investigate mutating both question and context together, so to create increasingly coherent Q&A samples.

References

[1] Xiang Li, Wei Zhang, Qian Ding, and Jian-Qiao Sun. Multi-layer domain adaptation method for rolling bearing fault diagnosis. *Signal processing*, 157:180–197, 2019.
[2] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mita- mura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.
[3] Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey. *arXiv preprint arXiv:2006.00632*, 2020.
[4] Egoitz Laparra, Steven Bethard, and Timothy A Miller. Rethinking domain adaptation for machine learning over clinical language. *JAMIA open*, 3(2):146–150, 2020.
[5]Garg, Siddhant and Goutham Ramakrishnan Bae: Bert-based adversarial examples for text classification. ArXiv preprint 2020.