Robust Q&BAE: Improving Out-of-Domain Question Answering Performance With Data Augmentation Techniques Inspired by Adversarial Perturbation Methods
By Lynn Kong, Phillip Weiss, and Adam Palhavani

Problem:
- Classical ML assumes training and test data come from the same distribution
  - Classification models trained on out-of-domain data can perform poorly
- Many prior works assume 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 approach on the task of the Robust QA track domain adaptation problem.

Background
- Several different categories of data distribution differ from the training distribution
- Our approach is closest to the supervised perturbation method that is in
- Several non-baseline models performed better than baseline (Table 1)

Results & Analysis
- Our approach is closest to the supervised perturbation method that is in
- Several non-baseline models performed better than baseline (Table 1)

Table 1. Validation performance of Phase 1 models.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Token Geo</th>
<th>Semi-Sense</th>
<th>PT</th>
<th>EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>48.286</td>
<td>32.481</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-MLM</td>
<td>54.728</td>
<td>38.696</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-MIBT</td>
<td>53.146</td>
<td>36.724</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-MIBT</td>
<td>51.764</td>
<td>37.612</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Methods & Experiments

Table 2. Overview of domain adaptation methods. The techniques studied here fall within the supervised domain adaptation / data-centric / rule-based categories and are evenly divided into perturbation methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data-centric</td>
<td>Unsupervised Domain Adaptation</td>
<td>Use labeled source data to adapt models to target domain</td>
</tr>
<tr>
<td>Rule-based</td>
<td>Supervised Domain Adaptation</td>
<td>Use labeled target data to adapt models to target domain</td>
</tr>
</tbody>
</table>

Training Procedure
- We tried ~14 combinations of language models, semantic similarity functions, number of masks, and number of indexes to consider. Each model takes ~4 hours, and there are >100 hyperparameters.
- We ran our experiments in two phases: first to test the simple version of the data-centric method and then to test the supervised version.

Phase 1:
- No supervision (Baseline), 14 combinations of language models, semantic similarity functions, number of masks, and number of indexes to consider.

Phase 2:
- No supervision (Baseline), 14 combinations of language models, semantic similarity functions, number of masks, and number of indexes to consider.

We hypothesize that since the 'relation extraction' training dataset contains several examples regarding genes and chromosomes, the BioBERT model performs better on this task.

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 QA.
- We observed that models trained on BERT-MIBT and BERT-MIBT perform better than baseline.

We demonstrate how the level of perturbation in the training set (number of mutations per sample and index upper bound) and found that mutation and ~7% of training data being perturbed samples performed best. However, this latter result is less conclusive.

In the future, we want to resolve the implementation limitations, such as creating a sliding window approach so that token changes can happen anywhere in the input text.

Additionally, we want to investigate mutating both question and context together, so to create increasingly coherent BQA samples.

References