Data Augmentation for Improved Generalizability of Natural Language Processing Models

Hee Hwang and Sudarshan Raghavan

April 22, 2021

Abstract

A rapid data augmentation framework to improve the performance of natural language processing models. To augment data for a particular downstream task, we use DepCC, A Dependency-Parsed Text Corpus from Common Crawl. First, we take the Common Crawl data and index 35M documents using Apache Solr, a search engine that uses BM25(Similar to TF-IDF) scoring model. Secondly, we prepare queries from the train/test dataset. After retrieving relevant documents using the query, we convert the documents into dense embeddings and apply K-nearest-neighbors to the candidate passages to rank the relevant documents. We augment data using various strategies. To show performance, we measure held-out accuracy.

1 Datasets

- ACL : Citation Intent Classification
- Hyper: HyperPartisan News Detection
- IMDb : Sentiment Classification

1.1 Size of Datasets

Task	train set	dev set	test set	# of Classes
ACL	1688	114	139	6
Hyper	516	64	65	2
IMDb	20000	5000	25000	2

1.2 Size of Augmented data

Task	Baseline	Strat. (i)	Strat. (ii)	Strat. (iii)
ACL	1688	11005	10248	10621
Hyper	516	1496	2184	2184
IMDb	20000	29492	67004	68666

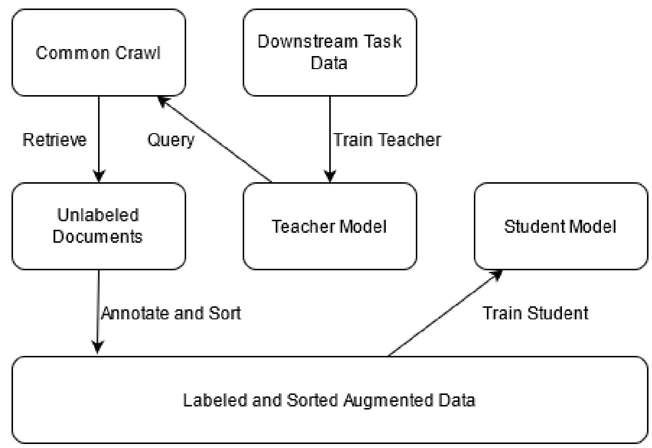
2 Results

Task	Baseline	Strat. (i)	Strat. (ii)	Strat. (iii)
ACL	62.5	60.5	64.4	67.1
Hyper	85.2	90.2	88.7	86.7
IMDb	93.8	92.4	91.9	92.3

3 Baseline models:

- An off-the-shelf RoBERTa model that has been finetuned to perform classification for each of the downstream tasks

4 Augmentation Model



5 Algorithm

```
1. Extract failed test examples from the baseline model
2. Retrieve passages/sentences from Common Crawl
3. Apply augmentation strategy (i)-(iii)
4. Augment all the labelled CC data to the training data
5. Retrieve RoBERTa on the augmented training set
```

6 Augmentation Strategies

- Strategy (i)
  Use baseline model (Teacher) to perform unsupervised labelling on retrieved CC data
- Strategy (ii)
  Using a task specific binary classifier, filter out retrieved CC data that is "out-domain"
  Use baseline model (Teacher) to perform unsupervised labelling on the filtered "in-domain" CC data
- Strategy (iii)
  Using a task specific binary classifier, filter out retrieved CC data that is "out-domain"
  Use ground truth labels of failed test examples and assign labels to the filtered "in-domain" CC data

7 TBD

Modify Query and Retrival / oversampling / downsampling
Perturb Query / Vary augmentation data / Measure Binary classifier