Data Augmentation Method for Fact Verification Using GPT-3

Jaehwan Jeong [1] Patrick Ryan [1,2] Harry Shin [1]

[1]Department of Computer Science [2]Department of Mathematics

Problem Description

Fact verification is a natural language processing task that tries to verify a claim with evidence. The problem can be broken down into two sub-tasks: information retrieval from a database and recognizing textual entailment (RTE). Currently, these datasets are expensive to create. Hence, it is desirable to develop a method to automatically and cheaply generate and label evidence-claim pairs. We propose a novel data augmentation method for Fact Verification RTE based on GPT-3.

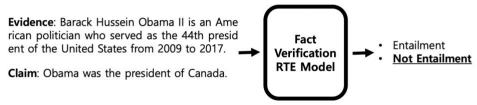

Figure 1. Sample RTE Task

Background

Prior research in machine-based data augmentation for fact verification has automatically generated evidence-claim pairs by fetching evidence from an external source, passing the evidence through a question generator to obtain a question-answer pair, and then creating a claim from the question-answer pair. Current proposed methods for generating non-entailing examples rely primarily on rule-based heuristics such as entity-swapping, which may be easily learned by a model and may not generalize to more challenging evidence-claim pairs.

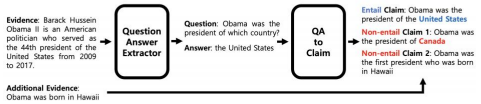

Figure 2. Previous machine-based data augmentation method

Lee et al. 2021 uses the GPT2-based perplexity score of a given evidence-claim pair to perform RTE in a few-shot manner. Their research is based on the hypothesis that a non-entailing sample will have a higher perplexity score than an entailing sample.

Type	Evidence	Claim	Perplexity
Entailment	Born in Saint James , Trinidad and Tobago and raised in South Jamaica , Queens , New York , Minaj earned public attention...	Nicki Minaj was born in Trinidad and Tobago.	3.82
Non-Entailment	Samsung Life Insurance is a South Korean multinational insurance company headquartered in Seoul...	Samsung Life Insurance is a multinational boy band.	55.85

Table 1. Perplexity Scores for FEVER samples

Methods

We propose a reverse approach from the method show in Figure 2, instead generating both evidence and claim from a question-answer pair obtained from an external source. To do this, we exploit GPT-3's pretrained knowledge to generate both evidence and claim based on an input question-answer pair. We observe that GPT-3 often incorporates its massive pretrained knowledge to generate evidence, often so much so that it does not preserve the semantic content of the original question-answer pair in the evidence. Hence, the evidence does not entail the claim.

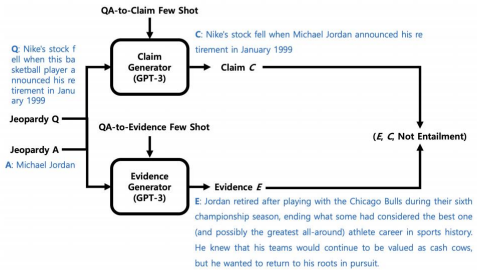

Figure 3. Proposed method for generating non-entailing evidence-claim pairs

We have empirically found that 90% of the generated evidence-claim pairs are non-entailing examples. We thus label all these examples as non-entailing and filter our samples with low perplexity scores that may potentially be entailing samples. Moreover, we also try generating a false claim by generating a false answer A' for a given QA pair. Given how GPT-3 almost always generates a true evidence, it will not entail a false claim based on Q and A'.

Datasets

As the baseline datasets, we use the Fact-Extraction and VERification (FEVER) dataset and Adversarial Natural Language Inference (ANLI) dataset. Then we augment the two datasets with 10,000 non-entailing samples generated using our proposed method. We use question-answer pairs from the Jeopardy! dataset as the seed data. To compare our augmentation method itself with a different augmentation method, we also generate 10,000 non-entailing samples using GPT-2 that has been fine-tuned on FEVER to generate a false claim given an evidence. Since ANLI maintains a non-entailing neutral class in addition to FEVER's entailment/contradictory split, we also maintained an extra version of the ANLI dev set, E/N, containing entailing and neutral samples to see if our data augmentation methods help distinguish such samples as well.

Class	FEVER	ANLI	GPT-2	GPT-3	FEVER DEV	ANLI DEV E/C	ANLI DEV E/N
Entailment	20,000	20,000	0	0	6,666	1,000	1,000
Non-Entailment	20,000	20,000	10,000	10,000	6,666	1,000	1,000
Total	40,000	40,000	10,000	10,000	13,332	2,000	2,000

Table 2. Dataset Distribution

Experiments and Results

We first fine-tune ELECTRA Small on the baseline FEVER and ANLI train sets and then observe how different augmentation methods improved model performance. We insert extra entailment samples from the original FEVER and ANLI dataset to the augmented train set to balance class.

Data	FEVER DEV	ANLI DEV E/C	ANLI DEV E/N
Baseline	0.924/0.923	0.552/0.571	0.612/0.608
Baseline+GPT2	**0.925**/**0.926**	0.566/0.611	0.623/0.645
Baseline+GPT3	**0.925**/**0.926**	**0.572**/**0.633**	**0.641**/ **0.672**

Table 3. Baseline vs Baseline with data augmentation (Accuracy / F1)

Here we compare the different filtering methods for our the GPT-3 augmented data. For High-PPL, we only keep 5,000 GPT-3 samples with highest perplexity scores, and for Mid-PPL, we keep the middle 5,000 samples. For False-Answer, we use 5,000 samples whose claims were generated based on question-false-answer pairs.

Data	FEVER DEV	ANLI DEV E/C	ANLI DEV E/N
Mid-PPL	0.923/0.923	0.560/0.611	0.622/0.644
High-PPL	**0.926**/ **0.926**	0.563/0.611	0.625/**0.648**
False-Answer	0.922/0.922	**0.572**/**0.613**	**0.629**/0.643

Table 4. Different filtering methods for GPT-3 generated samples (Accuracy / F1)

Analysis

- Dataset augmentation with GPT-2/GPT-3 improved accuracy and F1 score across all three dev sets, with pronounced improvement of 3% accuracy on the ANLI dev set
- Adding GPT-3 generated data of high perplexity yielded greater improvement on both accuracy and F1 score than adding data of average perplexity
- The boost in F1 score is mainly due to our augmentations improving recall while maintaining the same precision on predicting entailment
- Qualitatively, our data augmentation method on the ANLI train set can apparently incorporate information from multiple sentences to make accurate predictions despite a lack of explicit evidence of the claim. For example, our data augmented models were able to correctly classify the entailing evidence-claim pair "Evil under the sun is a video game released... the pc version was released in 2007, and the wii version one year later." and "The wii version came out in 2008" while the baseline model got it wrong. Furthermore, we observe the social and ethical impact of our research, with our data augmented models, for instance, correctly classifying the non-entailing evidence-claim pair "Passion play is a 2010 american drama film... executive produced by rebecca wang..." and "The executive producer was male."

Conclusions

We observe that the model performs better across all dev-sets when trained on data containing our automatically generated examples versus the baseline datasets. In particular, we observe that our method boosts model F1 accuracy on both ANLI `CONTRADICTION` and `NEUTRAL` by approximately 6% points.