GenAug: Data Augmentation For Finetuning Text Generators

Steven Y. Feng*1, Varun Gangal*1, Dongyeop Kang2, Teruko Mitamura1, Eduard Hovy1

1Language Technologies Institute, Carnegie Mellon University
2University of California, Berkeley

1st DeeLIIO Workshop, EMNLP 2020
Introduction

- GenAug: data augmentation for finetuning text generators
- Propose and evaluate various augmentation methods
- Investigate effects of the amount of augmentation
- Finetuning GPT-2 on a subset of Yelp Reviews
- Evaluate various aspects of the generated text
The Need for Augmentation for Generation

- Large pretrained generators like GPT-2 → Possibility to perform generation in many new domains and settings
  - In particular, low-resource domains with very little data
- GPT-2 still needs to be finetuned to the specific domain!
- Without this, it can’t pick up:
  - Length characteristics
  - Stylistic variables (e.g. formality, sentiment)
  - Domain-specific word choices
- Apart from specific tasks like MT, most augmentation methods in NLP have been focused on classification
Why Not Use Same Methods Directly?

- **Reason 1:** Generation is much more sensitive to the quality of “x”

- **In classification:** Maximize $P(y^*| x)$
  - Using augmentation: Maximize $P(y^*| x')$
  - Since $x'$ goes into conditional → More leeway for how noisy $x'$ can be.
  - Thinking of in model terms, as long as encoder representations shift only slightly, we can vary $x'$ quite a bit

- **In generation:** Maximize $\prod_i P(x_i | x_1, x_2, \ldots x_{i-1})$
  - $x'$ is both the target and the conditional
  - Affects loss and hence learning directly
Why Not Use Same Methods Directly?

- **Reason II:** Generation requires improving or maintaining performance on multiple metrics, not just the training loss
- **Fluency:** How fluent, grammatical, and natural is the text?
- **Diversity:** How diverse are the outputs given the same input?
- **Coherence:** Does the generated text maintain the same topic as the generation continues?
- Hence, methods that seemingly reduce training loss could still degrade other aspects of the text such as diversity
GPT-2 (Radford et al., NAACL ’19)

- OpenAI GPT-2 = Generatively Pretrained Transformers
- A left-to-right Transformer with 12 layers, ~117M parameters
- Pretrained on WebText → Corpus of newswire, forums, etc.
- Trained like a typical LM, maximize likelihood of word | left context
- Can be further fine-tuned by giving appropriately constructed text
- Conditional generation: complete discourse given prompt
Yelp Reviews Dataset

- Contains user reviews on businesses
- Substantially different in domain from GPT-2’s training data
- Contains long reviews that carry sentiment (1-5 star ratings)
- YLR: Randomly select a small subset, ~1%, for GenAug experiments
- Simple baseline: finetuning GPT-2 on YLR only
Augmentation Methods

- Suite of perturbation operations to generate augmented examples per original YLR review
- Motivated by intuition, greater focus on modestly meaning-altering perturbations, which toggle specific aspects of the text
- Synthetic Noise: character-level
- Synonym: word choice
- Hyponym/Hypernym: word granularity
- STE: topic-level semantics

<table>
<thead>
<tr>
<th>Method</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Review</td>
<td>got sick from the food. overpriced and the only decent thing was the bread pudding. wouldn't go back even if i was paid a million dollars to do so.</td>
</tr>
<tr>
<td>Synthetic Noise (10%)</td>
<td>got sick from the food. overpriced and the only decent thing was the scratch pud. wouldn't go back even if i was paid a one thousand thousand dollars to do so.</td>
</tr>
<tr>
<td>Synonym Replacement</td>
<td>got sick from the food. overpriced and the only decent thing was the crescent roll corn pudding. wouldn't go back even if i was paid a million kiribati dollar to do so.</td>
</tr>
<tr>
<td>Hyponym Replacement</td>
<td>got sick from the food. overpriced and the only decent thing was the baked goods dish. wouldn't go back even if i was paid a large integer dollars to do so.</td>
</tr>
<tr>
<td>Random Insertion (10%)</td>
<td>got sick from the food nauseous. overpriced and the only decent thing was the bread pudding. wouldn't go back even if i was paid a million dollars boodle to do so.</td>
</tr>
<tr>
<td>Semantic Text Exchange</td>
<td>got sick from the coffee. overpriced and the food was good. wouldn't come back if i was in a long hand washing machine.</td>
</tr>
</tbody>
</table>
Baseline: Random Trio

- Based on EDA: Easy Data Augmentation Techniques For Boosting Performance on Text Classification Tasks (Wei et al., EMNLP ’19)
  - Suite of simple, easy-to-implement random perturbation operations
  - Select one randomly each time to create augmented example
  - Tested on five classification tasks: SST-2, CR, SUBJ, TREC, Pro-Con

<table>
<thead>
<tr>
<th>Operation</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>A sad, superior human comedy played out on the back roads of life.</td>
</tr>
<tr>
<td>SR</td>
<td>A <em>lamentable</em>, superior human comedy played out on the <em>backward</em> road of life.</td>
</tr>
<tr>
<td>RI</td>
<td>A sad, superior human comedy played out on <em>funniness</em> the back roads of life.</td>
</tr>
<tr>
<td>RS</td>
<td>A sad, superior human comedy played out on <em>roads</em> back <em>the</em> of life.</td>
</tr>
<tr>
<td>RD</td>
<td>A sad, superior human out on the roads of life.</td>
</tr>
</tbody>
</table>

Table 1: Sentences generated using EDA. SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.
Baseline: Random Trio

- Easy Data Augmentation Techniques For Boosting Performance on Text Classification Tasks (Wei et al., EMNLP '19)
- Improvements observed on all five classification tasks

- Random Trio: take three of these perturbation operations for GenAug: random swap, random insertion, random deletion
Synthetic Noise

- Intuition: perturbations at the character-level shouldn’t perturb overall input representation
- Already happens in most corpora → e.g. spelling mistakes
- To more closely mimic humans, the first and last character of each word are left unperturbed
- Only perturb the prompt portions of reviews
- E.g. sick → seick, food → fotod
Keyword Replacement

- Replace keywords within YLR reviews with other words using WordNet

1. **Synonyms (WN-Syns):** Replace with a synonym of the same POS (e.g. *large* → *huge*)

2. **Hypernyms (WN-Hypers):** Replace with parent-word (of the same POS) from WordNet taxonomy (e.g. *dog* → *mammal*, *dollar* → *currency*)

3. **Hyponyms (WN-Hypos):** Replace with child-word (of the same POS) from WordNet taxonomy (e.g. *food* → *beverage*, *dragon* → *wyvem*)
Semantic Text Exchange (STE)

- New task proposed in *Keep Calm and Switch On! Preserving Sentiment and Fluency in Semantic Text Exchange* (Feng et al., EMNLP '19)

- **Example:**

<table>
<thead>
<tr>
<th>Original text:</th>
<th>This pepperoni pizza is great! The crust is filled with cheese and it comes with many toppings.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replacement entity:</td>
<td>sandwich</td>
</tr>
<tr>
<td>Desired output text:</td>
<td>This ham sandwich is great! The bread is filled with grains and it comes with many fillings.</td>
</tr>
</tbody>
</table>

- We use SMERTI: entity replacement, similarity masking, text infilling
- Entity to replace: noun keywords/phrases (to maintain diversity)
- Entity that replaces (RE): a noun keyword/phrase from training data
- Intuition: alter semantics of the entire text w.r.t. a particular topic
Augmentation Amounts

- Explore the effects of the amount of augmentation
- Test out 1.5x, 2x, 3x, and 4x augmentation
- E.g. 4x → each example has three augmentations
- Use combination of Synthetic Noise, STE, and keyword replacement
  - Each method augments 1/3 of YLR training examples
Evaluation: Text Fluency

Do the continuations sound like good, acceptable English?

1. **PPL (↓)**: Perplexity according to a language model M

   \[ PPL(S) = \exp(-\frac{1}{|S|} \ln(p_M(S))) \]

2. **SLOR (↑)**: PPL but normalizes for word frequency

   \[ SLOR(S) = \frac{1}{|S|}(\ln(p_M(S)) - \ln(\prod_{t \in S} p(t))) \]
Evaluation: Text Diversity

Are the continuations sufficiently non-repetitive? (Inter + Intra)

1. **SBLEU (↓):** The highest BLEU with one of the other continuations
   \[ E_{s \sim S}[\text{BLEU}(s, S - \{s\})] \]

2. **UTR (↑):** Ratio of unique to total trigrams, aggregated over all continuations

3. **TTR (↑):** Mean ratio of unique to total tokens per continuation
Evaluation: Semantic Content Preservation (SCP)

Do continuations have content related to the input prompt?

- **BPRO (↑):** Avg. BERTScore* between prompt and continuation
  - Measures strength of pairwise alignment between BERT embeddings of prompt and continuation

* As originally proposed in BERTScore: Evaluating Text Generation With BERT (Zhang et al., ICLR ’20)
Evaluation: Sentiment Consistency

- **SentStd (↓)**: Average standard deviation of sentiment among each batch of 100 continuations
  - Do all continuations per input prompt have similar sentiment?

- **SentDiff (↓)**: Mean abs. difference between sentiment of generated continuations (each concatenated with the input prompt) and ground-truth YLR reviews
  - Do continuations carry sentiment aligning with ground-truth text?
## Examples of Generated Outputs

<table>
<thead>
<tr>
<th>Method</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prompt</strong></td>
<td>i got my hair and make up done here for my wedding on 12 29 13 . everything was amazing . hannah styled my hair and the results were pure perfection . i</td>
</tr>
<tr>
<td><strong>Original</strong></td>
<td>wish my hair could look like that everyday . i only have positive things to say about this place and would definitely recommend this place . i loved everything about this place !</td>
</tr>
<tr>
<td><strong>Gold (Yelp-LR)</strong></td>
<td>went home feeling amazing . you get a full set that changes throughout the year . thanks so much again hannah ! you did an awesome job for me and my mom .</td>
</tr>
<tr>
<td><strong>Synthetic Noise</strong></td>
<td>am forever thankful for hannah and her store . she's been so nice and accommodating to my needs . she explained my wants and what i could do and she never backed off . i will definitely be back to her store . this is a terrific place for professional hair and make up</td>
</tr>
<tr>
<td><strong>WN-Hypers</strong></td>
<td>am so happy i came here and will absolutely continue coming here to get my perfect cut . i left well satisfied . i love this place! thanks yelpers and thank you hannah and make up artist anthony! you've earned my trust</td>
</tr>
<tr>
<td><strong>2x</strong></td>
<td>highly recommend this salon . they even have some coupons on their site . i also got my eyebrows and lip waxing here . very affordable too ! i'll be back for sure</td>
</tr>
<tr>
<td><strong>3x</strong></td>
<td>couldn't believe how beautifully my hair turned out . my stylist was very quick and made sure to check on my hair every step of the way . the environment is a bit loud , but the receptionists and staff make up for it with a great quality of service and product . the price is right for the quality of the work . you'll definitely want to check this place out . i can't wait to return</td>
</tr>
<tr>
<td><strong>4x</strong></td>
<td>have to say i will definitely return to this salon . it's very romantic and upscale , all of the staff is very friendly and welcoming . i would definitely recommend this place to anyone who wants a beautiful hairdresser</td>
</tr>
</tbody>
</table>

Table 4: Examples of generated continuations from GPT-2 finetuned on select augmentation methods & amounts. **Prompt** is the first half of the original Yelp review fed in as input, and **Original** is the ground-truth continuation.
Evaluation Results - Methods

- Two baselines:
  - Gold (Yelp-LR): finetuning without augmentation
  - Random Trio: three methods within EDA

- Synthetic Noise and WN-Hypernyms outperform on almost all metrics.

<table>
<thead>
<tr>
<th>Variations</th>
<th>Gold (Yelp-LR)</th>
<th>Random Trio</th>
<th>STE</th>
<th>Synthetic Noise</th>
<th>WN-Syns</th>
<th>WN-Hypos</th>
<th>WN-Hypers</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBLEU (↓)</td>
<td>0.2639</td>
<td>0.2727</td>
<td>0.2776</td>
<td>0.2572</td>
<td>0.2789</td>
<td>0.2691</td>
<td>0.2651</td>
</tr>
<tr>
<td>UTR (↑)</td>
<td>0.6716</td>
<td>0.6660</td>
<td>0.6495</td>
<td>0.6925</td>
<td>0.6540</td>
<td>0.6669</td>
<td>0.6808</td>
</tr>
<tr>
<td>TTR (↑)</td>
<td>0.7173</td>
<td>0.7176*</td>
<td>0.7056</td>
<td>0.7461</td>
<td>0.6978</td>
<td>0.7129</td>
<td>0.7296</td>
</tr>
<tr>
<td>RWords (↓)</td>
<td>-6.0637</td>
<td>-6.0718</td>
<td>-6.0508</td>
<td>-6.1105</td>
<td>-6.0801</td>
<td>-6.0895</td>
<td>-6.0841</td>
</tr>
<tr>
<td>SLOR (↑)</td>
<td>2.9377</td>
<td>2.9404*</td>
<td>2.8822</td>
<td>2.9851</td>
<td>2.9368*</td>
<td>2.9373*</td>
<td>2.9447</td>
</tr>
<tr>
<td>BPRO (↑)</td>
<td>0.0969</td>
<td>0.0994</td>
<td>0.0928</td>
<td>0.1022</td>
<td>0.0899</td>
<td>0.0925</td>
<td>0.1038</td>
</tr>
<tr>
<td>SentStd (↓)</td>
<td>0.0852</td>
<td>0.0836</td>
<td>0.0837</td>
<td>0.0821</td>
<td>0.0864</td>
<td>0.0859*</td>
<td>0.0827</td>
</tr>
<tr>
<td>SentDiff (↓)</td>
<td>0.0783</td>
<td>0.0773</td>
<td>0.0777*</td>
<td>0.0762</td>
<td>0.0782*</td>
<td>0.0793</td>
<td>0.0768</td>
</tr>
</tbody>
</table>

Table 2: Average results by variation. Bold values indicate results better than Gold (Yelp-LR). Arrows beside each metric indicate whether lower or higher is better. * indicates insignificant values (using an α of 0.05).
Methods: Fluency and SCP

SLOR (↑) by Variation

BPRO (↑) by Variation

Random Trio  STE  Synthetic Noise  WN-Syns  WN-Hypos  WN-Hypers

SLOR  Gold (Yelp-LR) SLOR

Random Trio  STE  Synthetic Noise  WN-Syns  WN-Hypos  WN-Hypers

BPRO  Gold (Yelp-LR) BPRO
Methods: Diversity

SBLEU (↓) by Variation

UTR (↑) & TTR (↑) by Variation
Methods: Sentiment Consistency

SentStd (↓) & SentDiff (↓) by Variation

- SentStd
- SentDiff

Random Trio, STE, Synthetic Noise, WN-Syns, WN-Hypos, WN-Hypers
Could Synthetic Noise be cheating its way to diversity by spuriously changing characters?

* Synthetic Noise would have more spelling errors than gold
* We run a spell-check on its outputs to assess this
  * **SpellWords (↓)**: Avg. # of misspelled words per continuation
  * **SpellChars (↓)**: Avg. # of character-level spelling mistakes per continuation
* Synthetic Noise actually reduces spelling errors

<table>
<thead>
<tr>
<th>Spellcheck</th>
<th>Gold (Yelp-LR)</th>
<th>Synthetic Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpellWords (↓)</td>
<td>3.0024</td>
<td>2.6274</td>
</tr>
<tr>
<td>SpellChars (↓)</td>
<td>4.5804</td>
<td>3.9190</td>
</tr>
</tbody>
</table>
Analysis (II) - Hypemysms vs. Hyponyms

- Why does WN-Hypers perform much better than WN-Hypos?
- Hyponyms sometimes introduce esoteric, rare words, which seldom occur apart from very specific contexts
  - E.g. dragon → wyvem, dollar → Kiribati dollar
- Unlike hyponyms, hypemym replacement maintains faithfulness to the original text. Example:
  - Hypemym: 3 dogs walked home. → 3 animals walked home.
  - Hyponym: 3 dogs walked home. → 3 Dalmatians walked home.
Analysis (III) - Semantic Text Exchange

- We perform STE using a sliding-window approach with 30-word windows: STE is performed on each and then concatenated.
- Each window contains a randomly selected RE.
- This may result in semantic inconsistencies between windows:
  - E.g. with REs "coffee" and "hand":

STE using SMERTI was also shown in Feng et al., 2019 to reduce fluency.
Analysis (IV) - Random Trio & WN-Syns

- Random Trio: random word-level noise seems to lead to poor generations and is much less suitable for GenAug.
- WN-Syns: synonym replacement likely does not adjust semantics of the text sufficiently and results in overfitting.
Amounts: Fluency and SCP

SLOR (↑) by Amount

BPRO (↑) by Amount
Amounts: Diversity

**SBLEU (↓) by Amount**

- 1.5x: 0.275
- 2x: 0.265
- 3x: 0.260
- 4x: 0.255

**UTR (↑) & TTR (↑) by Amount**

- 1.5x: 0.78
- 2x: 0.76
- 3x: 0.74
- 4x: 0.72
Amounts: Sentiment Consistency

SentStd (↑) & SentDiff (↓) by Amount

- SentStd
- SentDiff
- 1x (Yelp-LR) SentStd
- 1x (Yelp-LR) SentDiff
Conclusion

- We introduced **GenAug**: data augmentation for text generation, specifically finetuning text generators.
- We proposed a new suite of augmentation methods and evaluation metrics adapted for GenAug.
- Two methods: **Synthetic Noise** and **Keyword Replacement with Hypernyms** outperformed a random augmentation baseline and the no-augmentation case.
- Our augmentations improve quality of the generated text up to 3x the amount of original training data.
Thanks for Listening!

- Steven Y. Feng: syfeng@cs.cmu.edu
  Website: https://styfeng.github.io/
  Twitter: @stevenyfeng

- Varun Gangal: vgangal@cs.cmu.edu
  Website: https://www.linkedin.com/in/varungangal/
  Twitter: @VarunGangal

- Dongyeop Kang: dongyeopk@berkeley.edu
  Website: https://dykang.github.io/
  Twitter: @dongyeopkang

- Teruko Mitamura: teruko@cs.cmu.edu

- Eduard Hovy: hovy@cs.cmu.edu

<table>
<thead>
<tr>
<th>Method</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Review</td>
<td>got sick from the food, overpriced and the only decent thing was the bread pudding, wouldn't go back even if i was paid a million dollars to do so.</td>
</tr>
<tr>
<td>Synthetic Noise (10%)</td>
<td>got sick from the food, overpriced and the only decent thing was the bread pudding, wouldn't go back even if i was paid a million dollars to do so.</td>
</tr>
<tr>
<td>Synonym Replacement (3 keywords)</td>
<td>got sick from the food, overpriced and the only decent thing was the scratch pudding, wouldn't go back even if i was paid a one thousand dollars to do so.</td>
</tr>
<tr>
<td>Hyponym Replacement (3 keywords)</td>
<td>got sick from the food, overpriced and the only decent thing was the crescent roll corn pudding, wouldn't go back even if i was paid a kilibati dollar to do so.</td>
</tr>
<tr>
<td>Hypernym Replacement (3 keywords)</td>
<td>got sick from the food, overpriced and the only decent thing was the baked goods dish, wouldn't go back even if i was paid a large integer dollars to do so.</td>
</tr>
<tr>
<td>Random Insertion (10%)</td>
<td>got sick from the food, overpriced and the only decent thing was the bread pudding, wouldn't go back even if i was paid a million dollars noodle to do so.</td>
</tr>
<tr>
<td>Semantic Text Exchange (60% MRT)</td>
<td>got sick from the coffee, overpriced and the food was good, wouldn't come back if i was in a long hand washing machine.</td>
</tr>
</tbody>
</table>