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 DeeLIO 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
- Hypernym/Hyponym: word granularity
- STE: topic-level semantics

| Method                        | Text                                                                 |
|-------------------------------|----------------------------------------------------------------------|
| Original Review               | 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. |
| Synthetic Noise (10%)         | 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 million dollars to do so. |
| Synonym Replacement (3 keywords) | 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. |
| Hyponym Replacement (3 keywords) | 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. |
| Random Insertion (10%)        | 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 doodle to do so. |
| Semantic Text Exchange (60% MRT) | got sick from the coffee. overpriced and the food was good. wouldn't come back if i was in a long hand washing machine. |
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

| Operation | Sentence |
|-----------|----------|
| None      | A sad, superior human comedy played out on the back roads of life. |
| SR        | A *lamentable*, superior human comedy played out on the *backward* road of life. |
| RI        | A sad, superior human comedy played out on *funniness* the back roads of life. |
| RS        | A sad, superior human comedy played out on *roads* back the of life. |
| RD        | A sad, superior human out on the roads of life. |

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:

  | Original text:          | This *pepperoni pizza* is great! The *crust* is filled with *cheese* and it comes with many *toppings*. |
  |-------------------------|---------------------------------------------------------------------------------------------------|
  | Replacement entity:     | sandwich                                                                                         |
  | Desired output text:    | This *ham sandwich* is great! The *bread* is filled with *grains* and it comes with many *fillings*. |

- 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

   $$PPL(S) = \exp\left(-\frac{1}{|S|} \ln(p_M(S)) \right)$$

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

   \[ E_{s \sim S}[BLEU(s, S - \{s\})] \]

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

| Method          | Text                                                                                                                                 |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------|
| **Prompt**      | 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                                                                 |
| **Original**    | 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!                                                                 |
| **Gold (Yelp-LR)** | 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.                                                                 |
| **Synthetic Noise** | 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. |
| **WN-Hypers**   | 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. |
| 2x              | 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.                                                                 |
| 3x              | 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. |
| 4x              | 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. |

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-Hypemmys outperform on almost all metrics

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

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

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

SentStd
SentDiff
Gold (Yelp-LR) SentStd
Gold (Yelp-LR) SentDiff
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

| Spellcheck           | Gold (Yelp-LR) | Synthetic Noise |
|---------------------|----------------|-----------------|
| SpellWords (↓)      | 3.0024         | 2.6274          |
| SpellChars (↓)      | 4.5804         | 3.9190          |
Analysis (II) - Hypemymms vs. Hyponymms

- Why does WN-Hypers perform much better than WN-Hypos?
- Hyponymms sometimes introduce esoteric, rare words, which seldom occur apart from very specific contexts
  - E.g. dragon → wyvem, dollar → Kiribati dollar
- Unlike hyponymms, 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.
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.
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

**S BLEU (↓) by Amount**

- 1.5x
- 2x
- 3x
- 4x

**UTR (↑) & TTR (↑) by Amount**

- 1.5x
- 2x
- 3x
- 4x
Amounts: Sentiment Consistency

SentStd (↓) & SentDiff (↓) by Amount

- Blue: SentStd
- Orange: SentDiff
- Purple dashed line: 1x (Yelp-LR) SentStd
- Red dashed line: 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

| Method                        | Text                                                                 |
|-------------------------------|----------------------------------------------------------------------|
| Original Review               | 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. |
| Synthetic Noise (10%)         | got **sick** from the **fod**. **overpriced** and the only decent **ting** was the bread pudding. wouldn't go back even if i was paid a million dollars to do so. |
| Synonym Replacement (3 keywords) | 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. |
| Hyponym Replacement (3 keywords) | 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 **kiribati dollar** to do so. |
| Hypermym Replacement (3 keywords) | got sick from the food **nauseous**. 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. |
| Random Insertion (10%)        | 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. |
| Semantic Text Exchange (60% MRT) | got sick from the **coffee**. overpriced and the food was **good**. wouldn't **come back** if i was in a **long hand washing machine**. |