Augmentation

Tongshuang Wu[1] **Marco Tulio Ribeiro**[2] **Jeffrey Heer**[1] **Daniel S. Weld**[1]
[1]University of Washington [2]Microsoft Research
wtshuang@cs.uw.edu marcotcr@gmail.com {jheer,weld}@cs.uw.edu

Abstract

This document contains the instructions for preparing a manuscript for the proceedings of ACL 2020. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1 Introduction

2 Task Formalization & Modeling

In response to the objectives, we form the perturbation as a text generation task. Given a paired sentence x and x', We form the training prompts as and finetune GPT-2 model

We combined multiple existing NLP datasets in finetuning our GPT-2 perturbation model.

All datasets provide pairs of sentences (s_1, s_2). We first compute the control codes introduced in Table 1, based on the POS-tag and parsing structure heuristics. Then, to construct the training texts, we concatenate the two sentences, the control tag, and the special tokens.

To control *where to change*

3 Application 1: Data Collection

??

3.1 Annotation Procedure

3.2 Counterfactual Data Augmentation

4 Related Work

5 Discussion and Conclusion

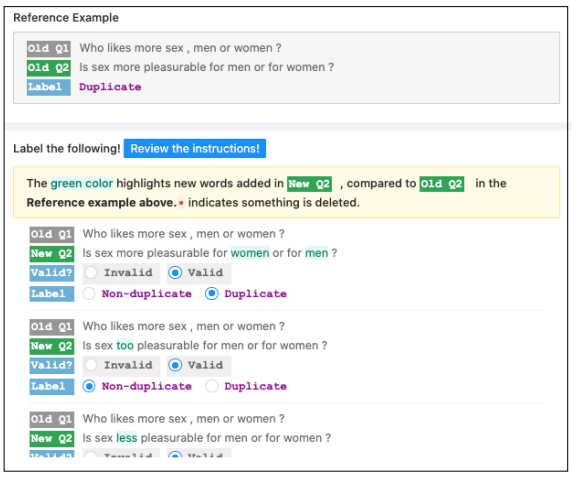

Figure 1: A sample labeling task. For each round of labeling, the annotator is given the original instance (and its label) as a reference, and they are tasked to label three variations of the instance by (1) grammatically validity and (2) classification task label. A more detailed instruction is in §B. [Sherry: This is placeholder screenshot. Change width/height ratio, choose a better example]

References

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. *Findings of EMNLP*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Daniel Khashabi, Tushar Khot, and A. Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. *EMNLP*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

Control code	Definitions and Examples	Datasets
`negation`	You'd figure that out → never know by watching it though.	(Kaushik et al., 2019; Gardner et al., 2020)
`quantifier`	Where can buy Jordan 5 → 6 shoes?	(Gardner et al., 2020)
`lexical`	*Changing just one word or noun chunks without breaking the POS tags.* He found them exciting → dull.	(Sakaguchi et al., 2019)
`resemantic`	*To replace short phrases or clauses without affecting the parsing tree.* How do you access Snapchat → brand yourself online?	(Wieting and Gimpel, 2017)
`insert`	*To add constraints without affecting the parsing structure of other parts.* I liked a Bangali boy.	(Wieting and Gimpel, 2017)
`delete`	*To remove constraints without affecting the parsing structure of other parts.* The lawyers paid the tourists.	(Wieting and Gimpel, 2017)
`restructure`	*To alter the dependency tree structure, e.g., changing from passive to positive.* How do you study well → animals?	(Zhang et al., 2019; McCoy et al., 2019)
`shuffle`	*To move (or swap) key phrases or entities around the sentence.* Why do so many more women → men commit suicide than men → women?	(Zhang et al., 2019; McCoy et al., 2019)

Table 1: A list of control codes used for semantically driving the GPT-2 generation, the model generated examples, and the training datasets that contains most of the corresponding patterns. [Sherry: Change all the examples to be on an identical sentence, not all different cases. And consider further annotate the tags based on whether they just do semantic change or also syntactic change.]

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

A Datasets for GPT-2 Finetuning

We combined the following NLP datasets in finetuning our GPT-2 perturbation model. To achieve a more balanced distribution, for each dataset, we extract control codes from all the data pairs available, and randomly sample up to 10k instances per control code. The distribution is shown in Table 2.

Contrast set In (Gardner et al., 2020), authors of 10 existing NLP dataset manually each perturbed 100-1,000 test instances in small but meaningful ways that change the gold label, so to inspect a model's decision boundary around a local instance. [Sherry: Make sure NLP is expanded somewhere in intro] The perturbation pattern varies based on the tasks and the annotators, allowing us to learn diverse perturbation methods. To make sure we can use the contrast set to evaluate the sentiment analysis model, we excluded the IMDB movie review from the training. [Sherry: Re-train the model with other contrast sets.]

Counterfactually-augmented data (CAD) To augment the training data, Kaushik et al. (2019) crowdsourced counterfactual perturbations for IMDB movie review (1.7k perturbations on 1.7k original instances) and SNLI (6.6k perturbations on 1.67k original instances). Similar to contrast set, the perturbation patterns vary based on the task, but can especially contribute to `negation`. We split the movie review paragraphs into paired sentences, to match the sentence length of other datasets.

WINOGRANDE is a large-scale dataset of 44k problems for testing common sense problems (Sakaguchi et al., 2019). The dataset contains nearly identical sentences that differ only by one trigger word, which flips the correct answer choice for certain questions. The dataset is most suitable for lexical tokens that are suitable for similar use cases.

ParaNMT-50M contains 50 million English-English sentential paraphrase pairs, covering various domains and styles of text, as well as different sentence structures Wieting and Gimpel (2017).

PAWS contains paraphrase and non-paraphrase pairs with high lexical overlap. Zhang et al. (2019) created 108k challenging pairs by controlled word swapping and back translation. As a result, the dataset demonstrates the `shuffle` and `restructure` strategy.

HANS is a controlled evaluation dataset designed for testing decision boundaries of NLI models (McCoy et al., 2019). The dataset contains 10k pairs of premises and hypotheses created based on 10 heavily on fallible syntactic heuristic rules, and therefore compensates rarer structural changes that may be missed by PAWS.

B Labeling Instructions & Quality

Procedure[1] The study started with an introduction, in which we explained the context and tasks (Figure 2): "given a reference example, the crowdworker should annotate its perturbed variations, based on whether the perturbation is valid (*sounds natural*), and the classification task label." To familiarize them with the task, we asked them to complete 1-2 training rounds, and explained the expected labels. The annotator then completed 22 tasks, each labeling 3 variations of a single example. The 22 rounds consisted of 20 actual labeling tasks and 2 extra "gold rounds" (6 labeled examples), with unambiguous examples and known groundtruth labels, which later served as filters for high quality crowdworkers. As a result, each annotator contributed $20 \times 3 = 60$ labels. The median task completion time was around 15-20 minutes (14.9 for QQP, 16.7 for Sentiment, and 19.8 for NLI), and participants received an average payment of $2.5 (equivalent to an hourly wage of $7.5).

Participants We recruited participants from Amazon's Mechanical Turk (MTurk), limiting the pool to subjects from within the United States with a prior task approval rating of at least 97% and a minimum of 1,000 approved tasks.

Data quality We applied two filtering strategies: (1) *High quality worker.* We only kept data from participants whose median labeling time was more than 18 seconds and correctly labeled at least 4 gold examples (out of 6), or who correctly labeled all gold ones. (2) *Majority vote labeling.* We collected two annotations per perturbation, and only kept data points that at least one annotator deems *valid*, and both annotators agree on a particular *class label.*

As such, when set out to collect augmentations on 1,000 original examples (thus 3,000 perturbations), we typically collect perturbations for 1,000

[1] A complete annotation demo is available at https://tongshuangwu.github.io/perturb-exp-ui/?assignmentId=1&platform=standalone&debug=debug&task=nli.

Dataset	negation	quantifier	leixcal	resemantic	insert	delete	restructure	shuffle	*global*
Contrast	147	194	652	743	349	354	707	95	459
CAD	3,654	431	7,299	3,832	2,401	2,403	3,798	822	3,702
WINOGRAND	3,817	110	10,000	7243	170	171	4,143	235	3,399
PAWS	75	280	5,672	2,730	4,161	4,155	10,000	10,000	10,000
PARANMT	2,477	507	5,196	10,000	5,436	5,424	10,000	1,594	10,000
HANS	50	0	0	0	3,893	3,891	798	1,139	229
Crawled	0	0	5,000	0	5,000	5,000	0	93	5,000
Total	10,220	1,522	33,819	24,548	21,407	21,398	29,446	13,978	32,789

Table 2: The datasets used for finetuning the GPT-2 perturbation model, and the control code distributions.

perturbations on 600 original examples. One of the authors labeled a subset of 100 perturbations on 100 original examples in Sentiment, and reached high agreement with the majority-voted results ($\kappa = 0.77$, the raw labeling agreement 88%).

Validity

Labeling efficiency Labeling three variations of a given example is reasonably easy for two reasons. First, instead of manually *generating* the perturbations, the annotators merely need to *verify* the machine-generated ones. Second, instead of having to parse the full sentence, they annotate based on the reference example and the corresponding perturbed phrases, a boost of efficiency also observed in (Khashabi et al., 2020). As a result, the median time for labeling one round (three perturbations) is 30 seconds, which is considerably shorter compared to existing manual perturbation methods: Kaushik et al. (2019)reported that workers spent roughly 5 minutes per revised IMDB movie review, and 4 minutes per revised sentence (for NLI). Similarly, Gardner et al. (2020) mentioned that three expert annotators spent 70 hours to create 588 counterfactual examples for IMDB movie review. Even for shorter image captions in NLVR2 visual reasoning dataset (Suhr et al., 2018), annotations would take approximately 30 seconds for one textual perturbation.

That said, manual annotations are undoubtedly more targeted. For example, the model is much less likely to flip an NLI instance from *contradiction* to *entailment*, if it is changing the hypothesis sentence regardless of the premise. We discuss the opportunity for interactive and more in-context perturbation in §5.

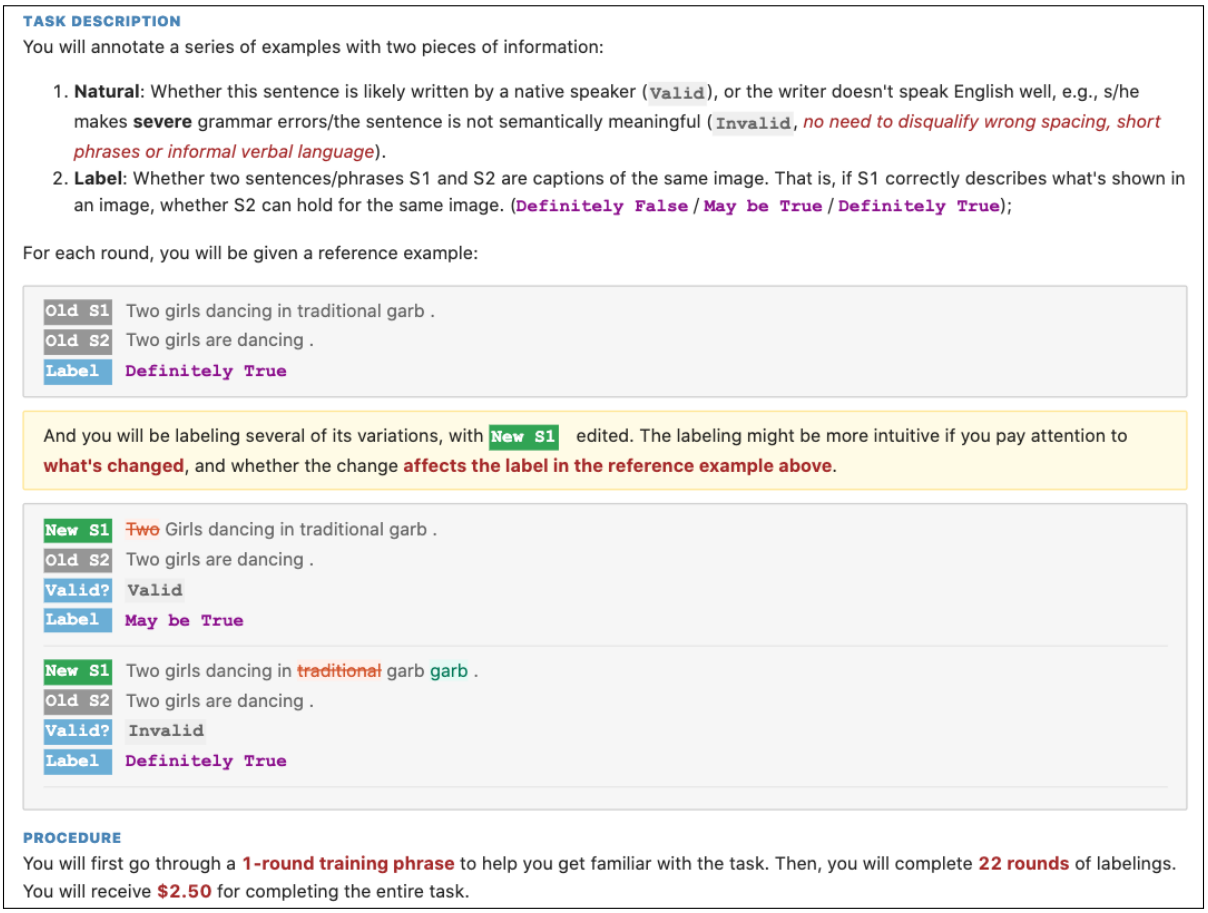

Figure 2: A sample instruction for the NLI task, with annotators providing labels based for the perturbed hypotheses (*New S2*). Instructions are similar for QQP and Sentiment, except for the label definitions and the examples. [Sherry: Change to hypothesis screenshot.]