An Empirical Survey of Data Augmentation for Limited Data Learning in NLP

Anonymous ACL submission

Abstract

NLP has achieved great progress in the past decade through the use of neural models and large labeled datasets. The dependence on abundant data prevents NLP models from being applied to low-resource settings or novel tasks where significant time, money, or expertise is required to label massive amounts of textual data. Recently, data augmentation methods have been explored as a means of improving data efficiency in NLP. To date, there has been no systematic empirical overview of data augmentation for NLP in the limited labeled data setting, making it difficult to understand which methods work in which settings. In this paper, we provide an empirical survey of recent progress on data augmentation for NLP in the limited labeled data setting, summarizing the landscape of methods (including token-level augmentations, sentence-level augmentations, adversarial augmentations and hidden-space augmentations) and carrying out experiments on 11 datasets covering topics/news classification, inference tasks, paraphrasing tasks, and single-sentence tasks. Based on the results, we draw several conclusions to help practitioners choose appropriate augmentations in different settings and discuss the current challenges and future directions for limited data learning in NLP.

1 Introduction

Deep learning methods have achieved strong performance on a wide range of supervised learning tasks (Sutskever et al., 2014; Deng et al., 2013; Minaee et al., 2021). Traditionally, these results were attained through the use of large, well-labeled datasets. This make them challenging to apply in settings where collecting a large amount of high-quality labeled data for training is expensive. Moreover, given the fast-changing nature of real-world applications, it is infeasible to relabel every example whenever new data comes in. This highlights a need for learning algorithms that can be trained with a limited amount of labeled data.

There has been a substantial amount of research towards learning with limited labeled data for various tasks in the NLP community. One common approach for mitigating the need for labeled data is **data augmentation**. Data augmentation (Feng et al., 2021) generates new data by modifying existing data points through transformations that are designed based on prior knowledge about the problem's structure (Yang, 2015; Wei and Zou, 2019). This augmented data can be generated from labeled data, and then directly used in supervised learning (Wei and Zou, 2019), or in semi-supervised learning for unlabeled data through consistency regularization (Xie et al., 2020) ("consistency training"). While various approaches have been proposed to tackle learning with limited labeled data — including unsupervised pre-training (Peters et al., 2018; Devlin et al., 2019; Raffel et al., 2020), multi-task learning (Glorot et al., 2011; Liu et al., 2017; Augenstein et al., 2018), semi-supervised learning (Zhu, 2005; Chapelle et al., 2009; Miyato et al., 2017; Xie et al., 2020), and few-shot learning (Deng et al., 2019) — in this work, we focus on and compare different data augmentation methods and their application to supervised and semi-supervised learning.

In this survey, we comprehensively review and perform experiments on recent data augmentation techniques developed for various NLP tasks. Our contributions are three-fold: (1) summarize and categorize recent methods in textual data augmentation; (2) compare different data augmentation methods through experiments with limited labeled data in supervised and semi-supervised settings on 11 NLP tasks, and (3) discuss current challenges and future directions of data augmentation, as well as learning with limited data in NLP more broadly. Our experimental results allow us to conclude that no single augmentation works best for every task, but (i) token-level augmentations work well for supervised learning, (ii) sentence-level augmentation usually works the best for semi-supervised learning, and (iii) augmentation methods can sometimes hurt performance, even in the semi-supervised setting.

Related Surveys. Recently, several surveys also explore the data augmentation techniques for NLP (Hedderich et al., 2020; Feng et al., 2021). Hedderich et al. (2020) provide a broad overview of techniques for NLP in low resource scenarios and briefly cover data augmentation as one of several techniques. In contrast, we focus on data augmentation and provide a more comprehensive review on recent data augmentation methods in this work. While Feng et al. (2021) also survey task-specific data augmentation approaches for NLP, our work summarizes recent data augmentation methods in a more fine-grained categorization. We also focus on their application to learning from limited data by providing an **empirical study** over different augmentation methods on various benchmark datasets in both supervised and semi-supervised settings, so as to hint data augmentation selections in future research.

2 Data Augmentation for NLP

Data augmentation increases both the amount (the number of data points) and the diversity (the variety of data) of a given dataset (Cubuk et al., 2019). Limited labeled data often leads to overfitting on the training set and data augmentation works to alleviate this issue by manipulating data either automatically or manually to create additional augmented data.Such techniques have been widely explored in the computer vision field, with methods like geometric/color space transformations (Simard et al., 2003; Krizhevsky et al., 2012; Taylor and Nitschke, 2018), mixup (Zhang et al., 2018), and random erasing (Zhong et al., 2020; DeVries and Taylor, 2017). Although the discrete nature of textual data and its complex syntactic and semantic structures make finding label-preserving transformation more difficult, there nevertheless exists a wide range of methods for augmenting text data that in practice preserve labels. In the following subsections, we describe four broad classes of data augmentation methods:

2.1 Token-Level Augmentation

Token-level augmentations manipulate words and phrases in a sentence to generate augmented text while ideally retaining the semantic meaning and labels of the original text.

Designed Replacement. Intuitively, the semantic meaning of a sentence remains unchanged if some of its tokens are replaced with other tokens that have the same meaning. A simple approach is to fetch synonyms as words for substitutions (Kolomiyets et al., 2011; Yang, 2015; Zhang et al., 2015a; Wei and Zou, 2019; Miao et al., 2020). The synonyms are discovered based on pre-defined dictionaries such as WordNet (Kolomiyets et al., 2011), or similarities in word embedding space (Yang, 2015). However, improvements from this technique are usually minimal (Kolomiyets et al., 2011) and in some cases, performance may even degrade (Zhang et al., 2015a). A major drawback stems from the lack of contextual information when fetching synonyms—especially for words with multiple meanings and few synonyms. To resolve this, language models (LMs) have been used to replace the sampled words given their context (Kolomiyets et al., 2011; Fadaee et al., 2017; Kobayashi, 2018; Kumar et al., 2020). Other work preserves the labels of the text by conditioning on the label when generating the LMs' predictions (Kobayashi, 2018; Wu et al., 2019a). In addition, different sampling strategies for word replacement have been explored. For example, instead of sampling one specific word from candidates by LMs, Gao et al. (2019) propose to compute a weighted average over embeddings of possible words predicted by LMs as the replaced input since the averaged representations could augment text with richer information.

Random Insertion, Replacement, Deletion and Swapping. While well-designed local modifications can preserve the syntax and semantic meaning of a sentence (Niu and Bansal, 2018), random local modifications such as deleting certain tokens (Iyyer et al., 2015; Wei and Zou, 2019; Miao et al., 2020), inserting random tokens (Wei and Zou, 2019; Miao et al., 2020), replacing non-important tokens with random tokens (Xie et al., 2017, 2020; Niu and Bansal, 2018) or randomly swapping tokens in one sentence (Artetxe et al., 2018; Lample et al., 2018; Wei and Zou, 2019; Miao et al., 2020) can preserve the meaning in practice. Different kinds of operations can be further combined (Wei

Methods	Level	Diversity	Tasks	Related Work
Synonym replacement	Token	Low	Text classification Sequence labeling	Kolomiyets et al. (2011), Zhang et al. (2015a), Yang (2015), Miao et al. (2020), Wei and Zou (2019)
Word replacement via LM	Token	Medium	Text classification Sequence labeling Machine translation	Kolomiyets et al. (2011), Gao et al. (2019) Kobayashi (2018), Wu et al. (2019a) Fadaee et al. (2017)
Random insertion, deletion, swapping	Token	Low	Text classification Sequence labeling Machine translation Dialogue generation	Iyyer et al. (2015), Xie et al. (2017) Artetxe et al. (2018), Lample et al. (2018) Xie et al. (2020), Wei and Zou (2019)
Compositional Augmentation	Token	High	Semantic Parsing Sequence labeling Language modeling Text generation	Jia and Liang (2016) , Andreas (2020) Nye et al. (2020), Feng et al. (2020) Furrer et al. (2020) , Guo et al. (2020)
Paraphrasing	Sentence	High	Text classification Machine translation Question answering Dialogue generation Text summarization	Yu et al. (2018), Xie et al. (2020) Chen et al. (2019), He et al. (2020) Chen et al. (2020c), Cai et al. (2020)
Conditional generation	Sentence	High	Text classification Question answering	Anaby-Tavor et al. (2020), Kumar et al. (2020) Zhang and Bansal (2019), Yang et al. (2020)
White-box attack	Token or Sentence	Medium	Text classification Sequence labeling Machine translation	Miyato et al. (2017), Ebrahimi et al. (2018b) Ebrahimi et al. (2018a), Cheng et al. (2019), Chen et al. (2020d)
Black-box attack	Token or Sentence	Medium	Text classification Sequence labeling Machine translation Textual entailment Dialogue generation Text Summarization	Jia and Liang (2017) Belinkov and Bisk (2017), Zhao et al. (2017) Ribeiro et al. (2018), McCoy et al. (2019) Min et al. (2020), Tan et al. (2020)
Hidden-space perturbation	Token or Sentence	High	Text classification Sequence labeling Speech recognition	Hsu et al. (2017), Hsu et al. (2018) Wu et al. (2019b), Chen et al. (2021) Malandrakis et al. (2019), Shen et al. (2020)
Interpolation	Token	High	Text classification Sequence labeling Machine translation	Miao et al. (2020), Chen et al. (2020c) Cheng et al. (2020b), Chen et al. (2020a) Guo et al. (2020)

Table 1: Overview of different data augmentation techniques in NLP. Diversity refers to the difference of augmented data from existing data and the amount of different augmented data could be generated.

and Zou, 2019), where each example is randomly augmented with one of insertion, deletion, and swapping. These noise-injection methods can efficiently be applied to training, and show improvements when they augment simple models trained on small training sets. However, the improvements might be unstable due to the possibility that random perturbations change the meanings of sentences (Niu and Bansal, 2018). Also, finetuning large pre-trained models on specific tasks might attenuate improvements due to preexisting generalization abilities of the model (Shleifer, 2019).

Compositional Augmentation. To increase the compositional generalization abilities of models, recent efforts have also focused on compositional augmentations (Jia and Liang, 2016; Andreas, 2020) where different fragments from different sentences are re-combined to create augmented examples. Compared to random swapping, compositional augmentation often requires more carefully-designed rules such as lexical overlap (Andreas, 2020), neural-symbolic stack machines (Chen et al., 2020e), and neural program synthesis (Nye et al., 2020). With the potential to greatly improve the generalization abilities to out-of-distribution data, compositional augmentation has been utilized in sequence labeling (Guo et al., 2020), semantic parsing (Andreas, 2020; Nye et al., 2020; Furrer et al., 2020), language modeling (Andreas, 2020; Shaw et al., 2020), and text generation (Feng et al., 2020).

2.2 Sentence-Level Augmentation

Instead of modifying tokens, sentence-level augmentation modifies the entire sentence at once.

Paraphrasing. Paraphrasing has been widely adopted as a data augmentation technique in various NLP tasks (Yu et al., 2018; Xie et al., 2020; Kumar et al., 2019; He et al., 2020; Chen et al., 2020b,c; Cai et al., 2020), as it generally provides more diverse augmented text with different word choices and sentence structures while preserving the meaning of the original text. The most popular is round-trip translation (Sennrich et al., 2015; Edunov et al., 2018), a pipeline which first translates sentences into certain intermediate languages and then translates them back to generate paraphrases. Translating through intermediate languages with different vocabulary and linguistic structures can generate useful paraphrases. To ensure the diversity of augmented data, sampling and noisy beam search can also be adopted during the decoding stage (Edunov et al., 2018). Other work focuses on directly training end-to-end models to generate paraphrases (Prakash et al., 2016), and further augments the decoding phase with syntactic information (Iyyer et al., 2018; Chen et al., 2019), latent variables (Gupta et al., 2017), and sub-modular objectives (Kumar et al., 2019).

Conditional Generation. Conditional generation methods generate additional text from a language model, conditioned on the label. After training the model to generate the original text given the label, the model can generate new text (Anaby-Tavor et al., 2020; Zhang and Bansal, 2019; Kumar et al., 2020; Yang et al., 2020). An extra filtering process is often used to ensure high-quality augmented data. For example, in text classification, Anaby-Tavor et al. (2020) first fine-tune GPT-2 (Radford et al., 2019) with the original examples prepended with their labels, and then generate augmented examples by feeding the fine-tuned model certain labels. Only confident examples as judged by a baseline classifier trained on the original data are kept. Similarly, new answers are generated on the basis of given questions in question answering and are filtered by customized metrics like question answering probability (Zhang and Bansal, 2019) and n-gram diversity (Yang et al., 2020). Generative models used in this setting have been based on conditional VAE (Bowman et al., 2016; Hu et al., 2017; Guu et al., 2017; Malandrakis et al., 2019), GAN (Iyyer et al., 2018; Xu et al., 2018) or pre-trained language models like GPT-2 (Anaby-Tavor et al., 2020; Kumar et al., 2020). Overall, these conditional generation methods can create novel and diverse data that might be unseen in the original dataset, but require significant training effort.

2.3 Adversarial Data Augmentation

Adversarial methods create augmented examples by adding adversarial perturbations to the original data, which dramatically influences the model's predictions and confidence without changing human judgements. These adversarial examples (Morris et al., 2020; Zeng et al., 2020) could be leveraged in adversarial training (Goodfellow et al., 2015) to increase neural models' robustness, and can also be utilized as data augmentation to increase the models' generalization ability (Miyato et al., 2017; Cheng et al., 2019).[1]

White-Box methods rely on model architecture and parameters being accessible and create adversarial examples directly using a model's gradients. Unlike image pixel values that are continuous, textual tokens are discrete and cannot be directly modified based on gradients. To this end, adversarial perturbations are added directly to token embeddings or sentence hidden representations (Miyato et al., 2017; Zhu et al., 2020; Jiang et al., 2019; Chen et al., 2020d) which creates "virtual adversarial examples". Other approaches vectorize modification operations as the difference of one-hot vectors (Ebrahimi et al., 2018b,a), or find real word neighbors in a model's hidden representations via its gradients (Cheng et al., 2019).

Black-Box methods are usually model-agnostic since they do not require information from a model or its parameters and usually focus on task-specific heuristics for creating adversarial examples. For example, by enumerating feasible substitutions on the basis of word similarity and language models, Ren et al. (2019) and Garg and Ramakrishnan (2020) select adversarial word replacements which severely influence the predictions from the text classification model. To attack reading comprehension systems, Jia and Liang (2017) and Wang and Bansal (2018) insert distracting but meaningless sentences at different locations in paragraphs and Ribeiro et al. (2018) leverage rule-based paraphrasing to pro-

[1] For more detailed discussion on textual adversarial examples, please refer to recent comprehensive surveys (Zhang et al., 2020b; Huq and Pervin, 2020; Goel et al., 2021).

duce semantically-equivalent adversarial examples. Likewise, for multi-hop question answering, Jiang and Bansal (2019) insert shortcut reasoning sentences and Trivedi et al. (2020) constructed disconnected reasoning example by removing certain supporting facts. For NLI, (Mitra et al., 2020) use VerbNet and other Semantic Role Labelling resources to generate pair of sentences containing same set of words but have different meaning. For machine translation, Belinkov and Bisk (2017) attacks character-based models by natural or synthesized typos and Tan et al. (2020) further adopt subword morphology level attacks. Similar attacks also help dialogue generation (Niu and Bansal, 2019) and text summarization (Cheng et al., 2020a; Fan et al., 2018). Other methods do not rely in editing input text directly; Iyyer et al. (2018) leverage round-trip translation to generate paraphrases in given syntactic templates and Zhao et al. (2017) search for adversarial examples in underlying semantic space with GANs (Goodfellow et al., 2014). Some of these heuristics could be further refined to obtain simple adversarial data augmentation approaches. For example, McCoy et al. (2019) craft adversarial examples for natural language inference using sophisticated templates which create lexical overlap between the premise and the hypothesis to fool the model. Min et al. (2020) proposes two simple yet effective adversarial transformations that reverse the position of subject and object or the position of premise and hypothesis.

2.4 Hidden-Space Augmentation

This line of work generates augmented data by manipulating the hidden representations through perturbations such as adding noise or performing interpolations with other data points. Hidden-space perturbations augment existing data by adding perturbations to the hidden representations of tokens (Miyato et al., 2017; Zhu et al., 2020; Jiang et al., 2019; Chen et al., 2020d; Shen et al., 2020; Chen et al., 2021) or sentences (Hsu et al., 2018; Wu et al., 2019b; Malandrakis et al., 2019).

Interpolation-Based Methods. Interpolation-based methods create new examples and labels by linear combinations of existing data-label pairs. Given two data-label pairs, virtual data-label pairs are created through linear interpolations of the pair of data points. Such interpolation-based methods can generate infinite augmented data in the "virtual vicinity" of the original data space, thus improving the generalization performance of models. Interpolation-based methods were first explored in computer vision (Zhang et al., 2018), and have more recently been generalized to the text domain (Miao et al., 2020; Chen et al., 2020c; Cheng et al., 2020b; Chen et al., 2020a) by performing interpolation between original data and token-level augmented data in the output space (Miao et al., 2020), between original data and adversarial data in embedding space (Cheng et al., 2020b), or between different training examples in general hidden space (Chen et al., 2020c). Different strategies to select samples to mix have also been explored (Chen et al., 2020a; Guo et al., 2020; Zhang et al., 2020a) such as k-nearest-neighbours (Chen et al., 2020a) or sentence composition (Guo et al., 2020).

We summarize the preceding overview of recent widely-used data augmentation methods in Table 1, characterizing them with respect to augmentation levels, the diversity of generated data, and their applicable tasks.

3 Consistency Training with DA

While data augmentation (DA) can be applied in the supervised setting to produce better results when only a small labeled training dataset is available, data augmentation is also commonly used in semi-supervised learning (SSL). SSL is an alternative approach for learning from limited data that provides a framework for taking advantage of unlabeled data. Specifically, SSL assumes that our training set comprises labeled examples in addition to unlabeled examples drawn from the same distribution. Currently, one of the most common methods for performing SSL with deep neural networks is "consistency regularization" (Bachman et al., 2014; Tarvainen and Valpola, 2017). Consistency regularization-based SSL (or "consistency training" for short) regularizes a model by enforcing that its output doesn't change significantly when the input is perturbed. In practice, the input is perturbed by applying data augmentation, and consistency is enforced through a loss term that measures the difference between the model's predictions on a clean input and a corresponding perturbed version of the same input.

Formally, let f_θ be a model with parameters θ, $f_{\hat{\theta}}$ be a fixed copy of the model where no gradients are allowed to flow, x_l be a labeled datapoint with label y, x_u be an unlabeled datapoint, and $\alpha(x)$ be a data augmentation method. Then, a typical loss

	Methods	Types	News Classification		Topic Classification	
			AG News	20 Newsgroup	Yahoo Answers	PubMed
	None	-	78.8(8.9)	65.2(4.8)	56.6(9.4)	63.7(6.1)/49.3(3.9)
Supervised	SR	Token	79.4(5.9)	66.1(2.5)	56.0(10.1)	62.4(5.7)/48.3(3.9)
	LM		76.8(5.1)	60.0(14.4)	56.2(8.4)	60.9(3.0)/47.4(2.5)
	RI		79.5(4.9)	66.6(0.6)	57.3(12.0)	63.7(4.2)/49.4(2.1)
	RD		79.6(5.0)	66.8(3.0)	58.0(8.3)	63.4(5.0)/49.3(1.5)
	RS		79.5(5.3)	64.8(10.8)	57.1(10.3)	63.8(7.4)/49.5(3.3)
	WR		79.7(2.0)	**67.5(4.2)**	**59.3(8.9)**	**64.9(4.9)/49.4(2.5)**
	RT	Sentence	**80.1(4.3)**	65.1(7.9)	57.1(9.6)	60.2(5.1)/46.3(6.4)
	ADV	Hidden	78.2 (5.3)	65.5(1.6)	53.8(4.89)	37.4(2.6)/19.9(10.6)
	Cutoff		79.3(5.0)	66.6(1.4)	57.3(9.3)	60.5(8.3)/46.6(9.4)
	Mixup		80.0 (6.52)	65.9(3.1)	57.8(4.19)	51.4(19.3)/39.8(3.2)
Semi Supervised	SR	Token	69.6(29.3)	65.7(1.8)	51.4(9.4)	59.3(5.9)/43.1(11.9)
	LM		68.5(13.7)	68.3(2.1)	53.2(6.3)	61.5(6.6)/46.4(4.4)
	RI		65.8(5.5)	66.7(1.1)	50.5(3.2)	61.4(11.3)/44.4(17.4)
	RD		73.2(14.0)	66.1(3.3)	51.5(7.5)	59.3(7.1)/46.0(3.8)
	RS		71.6(16.6)	65.0(2.0)	51.1(7.1)	64.2(12.1)/46.7(11.5)
	WR		74.1(12.3)	**69.3(2.5)**	55.6(5.9)	60.4(7.5)/43.7(14.2)
	RT	Sentence	82.1(8.2)	68.8(2.4)	59.8(3.9)	**64.3(1.2)/49.8(1.9)**
	ADV	Hidden	**82.3(2.33)**	66.8(5.9)	55.9(3.89)	62.2(10.8)/46.2(9.8)
	Cutoff		79.9(5.5)	67.9(0.8)	**60.1(1.0)**	62.7(9.0)/48.1(3.2)

Table 2: Topic Classification and News Classification results with 10 examples. We report the average results across 3 different random seeds with the 95% confidence interval and **bold** the best results.. For PubMed, we report the accuracy and F1 score.

function for consistency training is

$$\mathrm{CE}(f_\theta(x_l), y) + \lambda_u \mathrm{CE}(f_{\hat{\theta}}(x_u), f_\theta(\alpha(x_u)))$$

where CE is the cross entropy loss and λ_u is a tunable hyperparameter that determines the weight of the consistency regularization term. In practice, various other measures have been used to minimize the difference between $f_{\hat{\theta}}(x_u)$ and $f_\theta(\alpha(x_u))$, such as the KL divergence (Miyato et al., 2018; Xie et al., 2020) and the mean-squared error (Tarvainen and Valpola, 2017; Laine and Aila, 2017; Berthelot et al., 2019). Because gradients are not allowed to flow through the model when it was fed the clean unlabeled input x_u, this objective can be viewed as using the clean unlabeled datapoint to generate a synthetic target distribution for the augmented unlabeled datapoint. We refer the reader the Appendix for more details.

4 Empirical Experiments

4.1 Datasets and Experiment Setup

To provide a quantitative comparison of the DA methods we have surveyed, we experiment with 10 of the most commonly used and model-agnostic augmentation techniques from different levels in Table 1, including: (i) *Token-level augmentation*: **S**ynonym **R**eplacement (**SR**) (Kolomiyets et al., 2011; Yang, 2015), Word Replacement based on **L**anguage **M**odel (**LM** (Kumar et al., 2020), **R**andom **I**nsertion (**RI**) (Wei and Zou, 2019; Miao et al., 2020), **R**andom **D**eletion (**RD**) (Wei and Zou, 2019), **R**andom **S**wapping (**RS**) (Wei and Zou, 2019), and **W**ord **R**eplacement (**WR**) based on TF-IDF in *Vocabulary Set* (Xie et al., 2020); (ii) *Sentence-level augmentation*: **R**oundtrip **T**ranslation (**RT**) (Xie et al., 2020; Chen et al., 2020c); (iii) *Hidden-space Augmentation*: **Adv**ersarial training (**ADV**) (Goodfellow et al., 2015), **Cutoff** (Shen et al., 2020), and **Mixup** in the embedding space (Zhang et al., 2018). Most aforementioned techniques are not label-dependent (except mixup), thus can be applied directly to unlabeled data.

We test them on different types of benchmark datasets including: (i) news classification tasks including AG News (Zhang et al., 2015b) and 20 Newsgroup (Joachims, 1997); (ii) topic classification tasks including Yahoo Answers (Chang et al., 2008) and PubMed news classification ((Zhang

	Methods	Types	Inference			Paraphrase		Single Sentence	
			MNLI	QNLI	RTE	QQP	MRPC	SST-2	CoLA
	None	-	35.2(0.7)	51.8(7.0)	49.8(3.1)	63.9(9.1)	61.8(21.2)	60.5(13.1)	12.9(6.32)
Supervised	SR	Token	35.1(2.3)	51.4(7.2)	51.5(3.4)	61.3(9.7)	59.7(26.3)	62.1(17.4)	7.2(11.6)
Supervised	LM	Token	35.3(0.8)	51.0(8.0)	49.0(1.4)	62.4(11)	61.0(24.3)	62.8(9.8)	6.8(15.8)
Supervised	RI	Token	34.9(2.6)	51.5(8.4)	**51.5(1.4)**	60.6(10.9)	60.6(25.0)	63.3(12.2)	7.8(7.42)
Supervised	RD	Token	**35.5(2.1)**	51.1(8.4)	50.9(2.4)	62.4(11.3)	61.2(22.0)	59.7(18.4)	7.1(16.6)
Supervised	RS	Token	35.1(1.1)	51.5(7.0)	50.9(5.0)	**62.6(6.7)**	**63.2(22.5)**	61.2(10.8)	5.2(17.0)
Supervised	WR	Token	34.5(2.6)	**52.0(3.8)**	50.0(0.9)	60.6(10.2)	61.0(25.3)	61.8(12.5)	7.0(10.6)
Supervised	RT	Sentence	35.3(0.5)	51.1(9.6)	50.8(4.4)	60.5(17.8)	61.8(23.7)	62.0(1.99)	8.37(8.35)
Supervised	ADV	Hidden	33.3(4.7)	49.7(1.8)	48.3(12.1)	57.5(24.7)	61.5(21.5)	53.3(13.07)	1.37(4.66)
Supervised	Cutoff	Hidden	35.1(2.3)	51.4(8.3)	52.2(3.6)	62.6(8.8)	61.0(21.2)	**63.5(8.45)**	**12.4(9.58)**
Supervised	Mixup	Hidden	32.6(3.5)	49.9(1.4)	49.8(9.2)	63.0(0.3)	62.1(19.8)	62.3(12.3)	4.03(8.68)
Semi-Supervised	SR	Token	35.6(1.0)	52.1(4.5)	**52.9(5.4)**	53.5(10.7)	68.1(4.0)	61.8(37.9)	6.65(5.69)
Semi-Supervised	LM	Token	35.0(3.3)	52.5(4.2)	50.2(6.5)	47.9(34.1)	68.4(3.8)	57.3(14.2)	6.38(6.3)
Semi-Supervised	RI	Token	**35.8(1.7)**	52.1(4.1)	50.7(1.4)	59.6(5.1)	64.9(8.9)	58.3(14.8)	6.55(0.91)
Semi-Supervised	RD	Token	35.2(0.5)	52.1(5.2)	52.6(4.9)	56.1(16.0)	62.4(30.6)	55.7(16.4)	4.33(10.9)
Semi-Supervised	RS	Token	34.6(2.5)	52.1(6.2)	51.5(3.7)	49.8(7.9)	63.2(22.5)	55.2(15.3)	7.77(11.77)
Semi-Supervised	WR	Token	34.8(2.5)	52.1(4.1)	50.9(1.8)	51.8(16.0)	63.1(23.5)	54.8(13.8)	5.43(17.8)
Semi-Supervised	RT	Sentence	35.3(2.7)	**52.7(4.8)**	51.6(4.1)	**63.9(7.5)**	62.2(12.5)	**61.9(20.8)**	**11.6(14.5)**
Semi-Supervised	ADV	Hidden	36.2(8.9)	50.6(1.9)	50.9(6.8)	59.1(14.7)	63.9(9.1)	53.1(5.0)	7.64(25.1)
Semi-Supervised	Cutoff	Hidden	35.3(2.8)	52.5(4.3)	51.7(6.5)	62.9(9.9)	**68.6(4.4)**	54.3(9.8)	4.11(11.8)

Table 3: GLUE results with 10 labeled examples per class. We report the average results across 3 different random seeds with the 95% confidence interval and **bold** the best results.

et al., 2015b) (iii) inference tasks including MNLI, QNLI and RTE (Wang et al., 2018); (iv) similarity and paraphrase tasks including QQP and MRPC (Wang et al., 2018); and (v) single-sentence tasks including SST-2 and CoLA (Wang et al., 2018). For all datasets, we experiment with 10 labeled data points per class [2] in a supervised setup, and an additional 5000 unlabeled data points per class in the semi-supervised setup. The detailed experimental setup is described in the Appendix.

4.2 Results

News/Topic Classification Tasks. The results are shown in Table 2. We observe that in supervised settings, *token-level augmentations* work the best. Specifically, word replacement works well, getting the highest or second highest score every time; in the semi-supervised settings, *sentence level augmentations* (round-trip translation) works the best, getting the highest or second highest score every time. This makes sense since for many classification tasks, multiple words indicate the label, and so dropping several words will not affect the label.

Inference Tasks. As shown in Table 3, we observe that *token-level augmentations* work the best overall (e.g., random insertion, random deletion, and word replacement) for both supervised and semi-supervised settings. This is a bit surprising since the inference tasks usually heavily depend on several words, and changing these words can easily change the label for inferene tasks.

Similarity and Paraphrase Tasks. From Table 3, in the supervised settings, we observe that *token-level augmentations* (random swapping) achieve the best performances, while *hidden space augmentations* work well in semi-supervised settings, with cutoff performing the best on average. This makes sense since for paraphrasing tasks, augmenting the text usually consists of paraphrases, and so can easily change whether two texts are paraphrases of each other.

Single Sentence Tasks. Based on the single-sentence tasks results in Table 3, *hidden space augmentations* (cutoff) provides the biggest boost in performance in supervised settings, while in semi-supervised settings, *sentence level augmentations* (roundtrip translation) works best. We note most augmentation methods hurt performance on CoLA, a task for judging grammatical acceptability. This could be caused by the fact that most of augmentation methods try to preserve meaning and not grammatical correctness.

Overall, **no single augmentation works the best for every task in the supervised or semi-**

[2]The results for 100 labeled data points per class are shown in the Appendix.

supervised setting. However, several overall conclusions can be made: first, augmentation does not always improve performance, and can sometimes hurt performances, even in the semi-supervised setting. This suggests that we may need to design different augmentations for different tasks. Second, token-level augmentations (especially word replacement and random swapping) work well in general for supervised learning, especially when there is extremely limited labeled data. Third, round-trip translation usually works the best for semi-supervised learning, showing the most consistent gains. However, if the computation is limited, cutoff may be a better choice.

5 Other Limited Data Learning Methods

This work mainly focuses on data augmentation and semi-supervised learning (consistency regularization) in NLP; however, there are other orthogonal directions for tackling the problem of learning with limited data. For completeness, we summarize this related work below.

Low-Resourced Languages. Most languages lack large monolingual or parallel corpora, or sufficient manually-crafted linguistic resources for building statistical NLP applications (Garrette and Baldridge, 2013). Researchers have therefore developed a variety of methods for improving performance on low-resource languages, including cross-lingual transfer learning which transfers models from resource-rich to resource-poor languages (Do and Gaspers, 2019; Lee and Lee, 2019; Schuster et al., 2019), few/zero-shot learning (Johnson et al., 2017; Blissett and Ji, 2019; Pham et al., 2019; Abad et al., 2020) which uses only a few examples from the low-resource domain to adapt models trained in another domain, and polyglot learning (Cotterell and Heigold, 2017; Tsvetkov et al., 2016; Mulcaire et al., 2019; Lample and Conneau, 2019) which combines resource-rich and resource-poor learning using an universal language representation.

Few-shot Learning. Few-shot learning is a broad technique for dealing with tasks with less labeled data based on prior knowledge. Compared to semi-supervised learning which utilizes unlabeled data as additional information, few-shot learning leverages various kinds of prior knowledge such as pre-trained models or supervised data from other domains and modalities (Wang et al., 2020). While most work on few-shot focuses on computer vision, few-shot learning has recently seen increasing adoption in NLP (Han et al., 2018; Rios and Kavuluru, 2018; Hu et al., 2018; Herbelot and Baroni, 2017). To better leverage pre-trained models, PET (Schick and Schütze, 2021a,b) finetune models using masked language modeling by converting the text and label into a fluent sentence, outperforming GPT3 for few shot learning (Brown et al., 2020).

6 Discussion and Future Directions

In this work, we empirically surveyed data augmentation methods for limited-data learning in NLP and compared them on 11 different NLP tasks. Despite the success, there are still certain challenges that need to be tackled for improve their performance. This section highlights some of these challenges and discusses future research directions.

Theoretical Guarantees and Data Distribution Shift. Current data augmentation methods for text typically assume that they are label-preserving and will not change the data distribution. However, these assumptions are often not true in practice, which can result in noisy labels or a shift in the data distribution and consequently a decrease in performance or generalization (e.g., QQP in Table 3). Thus, providing theoretical guarantees that augmentations are label- and distribution-preserving under certain conditions would ensure the quality of augmented data and further accelerate the progress of this field.

Automatic Data Augmentation. Despite being effective, current data augmentation methods are generally manually-designed. Methods for automatically selecting the appropriate types of data augmentation still remain under-investigated. Although certain augmentation techniques have been shown effective for a particular task or dataset, they often do not transfer well to other datasets or tasks (Cubuk et al., 2019), as shown in Table 3. For example, paraphrasing works well for general text classification tasks, but may fail for some subtle scenarios like classifying bias because paraphrasing might change the label in this setting. Automatically learning data augmentation strategies or searching for an optimal augmentation policy for given datasets/tasks/models could enhance the generalizability of data augmentation techniques (Maharana and Bansal, 2020; Hu et al., 2019).

References

Alberto Abad, Peter Bell, Andrea Carmantini, and Steve Renais. 2020. Cross lingual transfer learning for zero-resource domain adaptation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7383–7390. AAAI Press.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Philip Bachman, Ouais Alsharif, and Doina Precup. 2014. Learning with pseudo-ensembles. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060.

Kevin Blissett and Heng Ji. 2019. Zero-shot cross-lingual name retrieval for low-resource languages. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 275–280, Hong Kong, China. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online. Association for Computational Linguistics.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 830–835. AAAI Press.

O. Chapelle, B. Scholkopf, and Eds A. Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. 2021. Hiddencut: Simple data augmentation for natural language understanding with better generalization. In *ACL*.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020a. Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1241–1251, Online. Association for Computational Linguistics.

Jiaao Chen, Yuwei Wu, and Diyi Yang. 2020b. Semi-supervised models via data augmentation for classifying interactive affective responses. In *AffCon@AAAI*.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020c. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. 2020d. SeqVAT: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8801–8811, Online. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984.

Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, and Denny Zhou. 2020e. Compositional generalization via neural-symbolic stack machines. *Advances in Neural Information Processing Systems*, 33.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020a. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3601–3608.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020b. AdvAug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970, Online. Association for Computational Linguistics.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123.

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Mike Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero. 2013. Recent advances in deep learning for speech research at microsoft. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Shumin Deng, Ningyu Zhang, Zhanlin Sun, Jiaoyan Chen, and Huajun Chen. 2019. When low resource nlp meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Terrance DeVries and Graham W. Taylor. 2017. Improved regularization of convolutional neural networks with cutout.

Quynh Do and Judith Gaspers. 2019. Cross-lingual transfer learning with data selection for large-scale spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1455–1460, Hong Kong, China. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018a. On adversarial examples for character-level neural machine translation.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018b. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Lisa Fan, Dong Yu, and L. Wang. 2018. Robust neural abstractive summarization systems and evaluation against adversarial information. *Interpretability and Robustness for Audio, Speech and Language Workshop at Neurips 2018*.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. Genaug: Data augmentation for finetuning text generators. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 29–42.

Steven Y. Feng, Varun Gangal, Jason Wei, Chandar Sarath, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Association for Computational Linguistics Findings*.

Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. 2020. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference of Machine Learning*.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *stat*, 1050:20.

Demi Guo, Yoon Kim, and Alexander Rush. 2020. Sequence-level mixed sample data augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5547–5552, Online. Association for Computational Linguistics.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2017. A deep generative framework for paraphrase generation. In *Association for the Advancement of Artificial Intelligence*.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2017. Generating sentences by editing prototypes. In *Transactions of the Association for Computational Linguistics, Volume 6*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *CoRR*, abs/2010.12309.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309, Copenhagen, Denmark. Association for Computational Linguistics.

Wei-Ning Hsu, Hao Tang, and James Glass. 2018. Unsupervised adaptation with interpretable disentangled representations for distant conversational speech recognition. *Interspeech 2018*.

Wei-Ning Hsu, Yu Zhang, and James Glass. 2017. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Zhiting Hu, Bowen Tan, Russ R Salakhutdinov, Tom M Mitchell, and Eric P Xing. 2019. Learning data manipulation for augmentation and weighting. *Advances in Neural Information Processing Systems*, 32:15764–15775.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aminul Huq and Mst. Tasnim Pervin. 2020. Adversarial attacks and defense on texts: A survey.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Thorsten Joachims. 1997. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, page 143–151, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276, Portland, Oregon, USA. Association for Computational Linguistics.

Alex Krizhevsky, I. Sutskever, and G. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25(2).

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Chia-Hsuan Lee and Hung-Yi Lee. 2019. Cross-lingual transfer learning for question answering. In *arXiv*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Adyasha Maharana and Mohit Bansal. 2020. Adversarial augmentation policy search for domain and cross-lingual generalization in reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou. 2019. Controlled text generation for data augmentation in intelligent artificial agents. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 90–98, Hong Kong. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*, pages 617–628.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning based text classification: A comprehensive review. In *ACM Computing Survey*.

Arindam Mitra, Ishan Shrivastava, and Chitta Baral. 2020. Enhancing natural language inference using new and expanded training data sets and new learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. *International Conference on Learning Representations (ICLR)*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Low-resource parsing with crosslingual contextualized representations. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 304–315, Hong Kong, China. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Adversarial over-sensitivity and over-stability strategies for dialogue models. In *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Tong Niu and Mohit Bansal. 2019. Automatically learning data augmentation policies for dialogue tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1317–1323, Hong Kong, China. Association for Computational Linguistics.

Maxwell I. Nye, Armando Solar-Lezama, Joshua B. Tenenbaum, and Brenden M. Lake. 2020. Learning compositional rules via neural program synthesis.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual lstm networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency.

In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142, Brussels, Belgium. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *Computer Science*.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2020. Compositional generalization and natural language variation: Can a semantic parsing approach handle both?

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Sam Shleifer. 2019. Low resource text classification with ulmfit and backtranslation.

Patrice Y. Simard, David Steinkraus, and John C. Platt. 2003. Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 3104–3112.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204.

Luke Taylor and Geoff Nitschke. 2018. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples. *ACM Computing Surveys*, 53(3):1–34.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 575–581, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019a. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Zhanghao Wu, Shuai Wang, Yanmin Qian, and Kai Yu. 2019b. Data Augmentation Using Variational Autoencoder for Embedding Based Speaker Verification. In *Proc. Interspeech 2019*, pages 1163–1167.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. 2017. Data noising as smoothing in neural network language models. *CoRR*, abs/1703.02573.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Dp-gan: Diversity-promoting generative adversarial network for generating informative and diversified text. In *CoRR*.

Diyi Yang, William Yang Wang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *CoRR*, abs/1804.09541.

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2020. Openattack: An open-source textual adversarial attack toolkit.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020a. SeqMix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8566–8579, Online. Association for Computational Linguistics.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2495–2509, Hong Kong, China. Association for Computational Linguistics.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020b. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28:649–657.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples. In *International Conference on Learning Representations*.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *ICLR*.

Xiaojin Jerry Zhu. 2005. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

A Consistency Training with DA

Xie et al. (2020) showed that consistency training can be effectively applied to semi-supervised learning for NLP. To achieve stronger results, they introduce several other tricks including confidence thresholding, training signal annealing, and entropy minimization. Confidence thresholding applies the unsupervised loss only when the model assigns a class probability above a pre-defined threshold. Training signal annealing prevents the model from overfitting on easy examples by applying the supervised loss only when the model is less confident about predictions. Entropy minimization trains the model to output low-entropy (highly-confident) predictions when fed unlabeled data. We refer the reader to (Xie et al., 2020) for more details on these tricks.

B Experimental Setup

We use $BERT_{base}$ (Devlin et al., 2019) as the base language model and use the same hyperparameters across all datasets/methods. We utilize accuracy as the evaluation metric for all datasets except for CoLA (which uses Matthews correlation) and PubMed (which uses accuracy and Macro-F1 score). Because the performance can be heavily dependent on the specific datapoints chosen (Sohn et al., 2020), for each dataset, we sample labeled data from the original dataset with 3 different seeds to form different training sets, and report the average result. For every setup, we finetune the model with the same seed as the dataset seed (in contrast to many works which report the max across different seeds).

We train our models on NVIDIA 2080ti and NVIDIA V-100 gpus. Supervised experiments take 20 minutes, and semi-supervised experiments take two hours. The BERT-base model has 100M parameters. We use the same hyperaparameter across all datasets, and so only use the validation set to find the best model checkpoint. We use a learning rate of $2e^{-5}$, batch size of 16, ratio of unlabeled to labeled data of 3, and dropout ratio of 0.1 for different augmentation methods.

C Results for 100 Labeled Data per Class

News/Topic Classification Tasks The results are shown in Table 4. We observe that overall, in both the supervised settings and semi-supervised setting, all the methods perofrmly similarly, with 2 points of each other. This indicates that data augmentation methods work well with limited labeled data, and with more labeled data, its effectiveness is removed.

Inference Tasks As shown in Table 5, we observe that most augmentation methods hurt the performance in both the supervised and semi-supervised setting, with a greater drop in performance in the semi-supervised setting.

Similarity and Paraphrase Tasks Similar to *inference tasks*, we observe in Table 5 that most augmentation methods hurt the performance in both the supervised and semi-supervised setting, with a greater drop in performance in the semi-supervised setting.

Single Sentence Tasks Unlike *inference tasks* and *paraphrase tasks*, augmentations methods help performance, as seen in Table 5, except for CoLA. We hypothesize the reason is because most augmentatiom methods seek to preserves meaning, not grammatical correctness, which is what CoLA measures. In the supervised and semi-supervised setting, hidden level augmentations work well, with cutoff performing the best.

D Case Study

We analyze several data augmentation methods and check whether the label is preserved for these and if this affects its performance. We look at 25 examples for the best performing data augmentation method and the worst performing data augmentation method for *20 News Group* and *RTE*. For *20 News Group*, *Random Deletion* was the best performing, and *Language Model* was the worst performing. In both cases, there were no examples where the label flipped, which makes sense since the input is usually several paragraphs with multiple references to the topic. Several examples are shown in Appendix. For *RTE*, *Language Model* was the worst performing and *Cutoff* was the best performing augmentation. *Language Model* flipped 24% of the labels with 4% uncertain, while *Cutoff* flipped 4% of the labels

	Methods	Types	News Classification		Topic Classification	
		-	AG News	20 Newsgroup	Yahoo Answers	PubMed
Supervised	None	-	87.9(1.05)	79.5(0.3)	68.6(0.71)	75.2(1.5)/59.5(2.0)
	SR	Token	88.5(0.87)	80.0(2.2)	**69.7(1.62)**	76.5(1.0)/60.7(0.7)
	LM	Token	88.1(1.00)	**80.5(1.8)**	68.8(3.2)	75.8(2.5)/59.9(1.7)
	RI	Token	88.0(2.08)	80.1(3.1)	69.1(1.68)	76.2(2.9)/60.3(1.7)
	RD	Token	88.1(0.84)	80.2(2.9)	68.7(2.2)	**76.9(0.6)/60.9(0.6)**
	RS	Token	88.4(0.97)	79.5(2.1)	69.0(2.03)	76.6(0.2)/60.6(0.7)
	WR	Token	87.9(1.19)	79.3(2.5)	69.4(5.89)	76.4(1.8)/60.4(1.6)
	RT	Sentence	88.3(0.17)	80.4(0.7)	68.8(1.88)	76.1(0.5)/60.3(0.5)
	ADV	Hidden	87.6(0.33)	78.5(1.4)	67.4(0.74)	75.6(4.0)/59.8(3.5)
	Cutoff	Hidden	88.3(0.38)	79.8(1.0)	68.7(0.47)	75.9(1.3)/60.1(0.7)
	Mixup	Hidden	**88.6(1.31)**	80.5(3.4)	68.27(1.76)	74.8(1.8)/59.2(0.2)
Semi Supervised	SR	Token	88.8(0.95)	81.2(8.4)	68.8(1.3)	76.6(1.5)/60.7(1.8)
	LM	Token	88.4(1.87)	81.4(1.0)	68.8(1.8)	76.4(1.3)/60.4(0.7)
	RI	Token	88.4(1.45)	80.3(3.0)	68.4(2.64)	76.8(1.2)/60.7(1.1)
	RD	Token	88.7(0.5)	80.5(0.8)	68.8(1.66)	**77.1(1.0)/61.2(1.5)**
	RS	Token	88.5(1.35)	80.9(2.2)	68.7(1.67)	76.9(1.7)/61.0(1.5)
	WR	Token	87.7(1.35)	81.5(1.3)	68.7(1.2)	76.5(0.5)/60.6(1.0)
	RT	Sentence	88.7(0.40)	**81.7(1.0)**	**69.7(1.06)**	77.0(1.2)/61.6(1.1)
	ADV	Hidden	88.0(1.04)	80.4(2.9)	68.9(1.74)	76.7(1.5)/60.9(1.2)
	Cutoff	Hidden	**88.9(0.25)**	81.3(4.6)	69.3(1.76)	76.7(2.1)/60.7(3.1)

Table 4: Topic Classification and News Classification results with 100 examples. We report the average results across 3 different random seeds with the 95% confidence interval and **bold** the best results.. For PubMed, we report the accuracy and F1 score.

with 12% uncertain. We show several examples of when the label flipped for RTE in the Table 6.

	Methods	Types	Inference			Paraphrase		Single Sentence	
			MNLI	**QNLI**	**RTE**	**QQP**	**MRPC**	**SST-2**	**CoLA**
Supervised	None	-	45.0(6.9)	63.2(10.7)	59.9(3.1)	71.0(2.6)	68.1(7.4)	82.7(4.0)	28.7(9.5)
	SR	Token	44.6(7.2)	62.9(9.4)	61.0(10.0)	68.9(2.2)	66.7(4.4)	84.0(1.9)	24.6(5.1)
	LM	Token	45.4(6.2)	60.6(7.7)	**61.5(9.1)**	69.6(1.7)	67.2(2.8)	83.8(3.1)	18.5(9.7)
	RI	Token	**45.8(7.5)**	**64.2(10.7)**	60.0(11.3)	69.2(0.6)	69.1(4.8)	84.3(1.4)	27.3(19.9)
	RD	Token	43.7(8.4)	63.6(9.4)	59.2(9.0)	69.2(1.5)	69.2(5.5)	82.3(2.05)	20.2(21.5)
	RS	Token	42.4(6.2)	63.3(9.1)	57.8(11.9)	68.3(1.6)	69.0(3.4)	82.5(5.0)	24.3(20.8)
	WR	Token	44.6(6.3)	61.6(8.8)	57.8(9.3)	66.7(1.8)	66.9(6.4)	83.5(1.9)	17.7(23.3)
	RT	Sentence	44.8(7.8)	59.0(7.6)	60.4(5.7)	**69.9(4.0)**	**69.6(1.6)**	84.3(3.27)	19.2(7.63)
	ADV	Hidden	39.1(10.9)	50.1(3.1)	57.3(8.7)	63.7(1.9)	68.7 (6.3)	69.8(5.3)	16.5(9.2)
	Cutoff	Hidden	44.9(5.5)	63.0(10.2)	59.3(8.8)	69.9(0.7)	66.5(1.3)	**84.7(0.9)**	**26.0(16.3)**
	Mixup	Hidden	35.7(7.3)	51.4(4.4)	60.5(6.52)	64.5(5.4)	67.9 (7.1)	83.5(3.4)	20.1(18.8)
Semi-Supervised	SR	Token	42.9(7.3)	60.1(6.2)	58.5(9.7)	65.0(6.0)	67.6(3.1)	85.1(3.5)	18.9(6.7)
	LM	Token	43.7(4.5)	60.9(10.4)	56.9(8.3)	59.3(12.0)	70.0(4.4)	83.9(4.1)	21.7(6.8)
	RI	Token	44.7(4.6)	62.5(10.5)	56.0(6.3)	68.3(0.1)	67.0(3.9)	84.2(3.0)	23.0(10.3)
	RD	Token	41.4(2.9)	59.4(6.4)	56(0.0)	**69.3(2.8)**	70.4(7.4)	83.6(2.3)	13.1(6.1)
	RS	Token	40.3(2.0)	60.3(8.7)	56.4(11.6)	66.8(2.3)	69.0(3.4)	84.5(3.6)	19.4(2.7)
	WR	Token	43.9(3.1)	60.5(8.8)	56.3(7.1)	65.4(4.3)	67.2(2.1)	83.3(4.5)	16.9(6.2)
	RT	Sentence	**45.4(7.7)**	**63.8(5.0)**	**59.9(9.1)**	68.3(2.9)	67.5(0.7)	**83.9(1.7)**	20.4(3.6)
	ADV	Hidden	44.1(3.4)	58.1(4.0)	58.6(5.2)	63.0(10.8)	67.6(5.2)	80.0(7.3)	13.5(7.8)
	Cutoff	Hidden	42.7(4.2)	60.3(7.4)	57.9(12.6)	67.2(4.4)	**71.4(2.0)**	82.5(5.4)	**23.9(2.7)**

Table 5: GLUE results with 100 labeled examples per class. We report the average results across 3 different random seeds with the 95% confidence interval and **bold** the best results.

Original	**Cutoff (Best)**	**Language Model (Worst)**
Sentence 1: The Walt Disney Co. donated one of the world's most significant private collections of African artwork, yesterday, to the Smithsonian's National Museum of African Art. Sentence 2: Disney gave the Smithsonian a trove of sought-after African art.	Sentence 1: The Walt Disney Co. donated one of the world's most significant private collections of African artwork, yesterday, to the Smithsonian's National Museum of African one Sentence 2: Disney gave the Smithsonian a trove of south African art.	Sentence 1: The Walt Disney Co. donated one of the world's most significant private collections of African artwork [PAD] [PAD] [PAD] to the Smithsonian's National Museum of African Art. Sentence 2: Disney gave the Smithsonian a trove of [PAD] African art.
Entailment	Entailment	Not Entailment
Sentence 1: An explosion, followed by a raging fire, demolished a plastics factory, killing at least three people and injuring at least 37. Sentence 2: A massive blast at a plastics factory killed at least two people.	Sentence 1: An explosion, followed by a raging fire, demolished a the factory, killing at least three people and injuring at least 37. Sentence 2: A massive blast at a plastics factory killed at shot two people.	Sentence 1: An explosion, followed by [PAD] [PAD] fire, demolished a plastics factory, killing at least three people and injuring at least 37. Sentence 2: A massive blast at a plastics [PAD] killed at least two people.
Entailment	Entailment	Not Entailment
Sentence 1: The prize is named after Alfred Nobel, a pacifist and entrepreneur who invented dynamite in 1866. Nobel left much of his wealth to establish the award, which has honoured achievements in physics, chemistry, medicine, literature and efforts to promote peace since 1901.	Setence 1: The prize is named after Alfred Nobel, a pacifist and entrepreneur who invented dynamite in 1866. Nobel left much of his wealth to establish the nobel which has honoured achievements in physics, chemistry, medicine, literature and efforts to promote peace since 1901.	The prize is named after Alfred Nobel, a pacifist and entrepreneur who invented dynamite in 1866 . Nobel left much of his wealth [PAD] [PAD] [PAD] [PAD], which has honoured achievements in physics, chemistry, medicine, literature and efforts to promote peace since 1901.
Sentence 2: Alfred Nobel invented dynamite in 1866.	Sentence 2: Alfred Nobel invented dynamite in 1866.	Sentence 2: Alfred Nobel invented dynamite in 1866.
Entailment	Entailment	Not Entailment

Table 6: Examples of different data augmentation methods on RTE and whether they preserve the original label or not