Collecting Entailment Data for Pretraining:
New Protocols and Negative Results

Samuel R. Bowman[*]

Center for Data Science and Department of Linguistics, New York University
`bowman@nyu.edu`

Jennimaria Palomaki, Livio Baldini Soares, Emily Pitler
Google Research
`{jpalomaki,liviobs,epitler}@google.com`

Abstract

Textual entailment (or NLI) data has proven useful as *pretraining* data for tasks requiring language understanding, even when building on an already-pretrained model like RoBERTa. The standard protocol for collecting NLI was not designed for the creation of pretraining data, and it is likely far from ideal for this purpose. With this application in mind, we propose four alternative protocols, each aimed at improving either the ease with which annotators can produce sound training examples or the quality and diversity of those examples. Using these alternatives and a simple MNLI-based baseline, we collect and compare five new 8.5k-example training sets. Our primary results are solidly negative, with our baseline MNLI-style dataset yielding good transfer performance, but none of our four new methods (nor the recent ANLI) showing any improvements on that baseline. However, we do observe that all four of these interventions, especially the use of *seed sentences* for inspiration, reduce previously observed issues with annotation artifacts.

1 Introduction

The task of natural language inference (NLI; also known as textual entailment) has been widely used as an evaluation task when developing new methods for language understanding tasks, but it has recently become clear that high-quality NLI data can be useful in transfer learning as well. Several recent papers have shown that training large neural network models on natural language inference data, then fine-tuning them for other language understanding tasks often yields substantially better results on those target tasks (Conneau et al., 2017; Subramanian et al., 2018). This result holds even when starting from large models like BERT (Devlin et al., 2019) that have *already* been pretrained

extensively on unlabeled data (Phang et al., 2018; Clark et al., 2019; Liu et al., 2019b).

The largest general-purpose corpus for NLI, and the one that has proven most successful in this setting, is the Multi-Genre NLI Corpus (MNLI Williams et al., 2018). MNLI was designed for use in a benchmark task, rather than as a resource for use in transfer learning, and it was not developed on the basis of any kind of deliberate experimentation. Further, data collected under MNLI's data collection protocol has known issues with *annotation artifacts* which make it possible to perform much better than chance using only one of the sentences in each pair (Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018).

This work begins to ask what would be involved in collecting a similar dataset that is explicitly designed with transfer learning in mind. In particular, we consider four potential changes to the original MNLI data collection protocol that are designed to improve either the ease with which annotators can produce sound examples, or the quality and diversity of those examples, and evaluate their effects on transfer. We collect a baseline dataset of 8500 examples that follows the MNLI protocol with our annotator pool, followed by four additional datasets of the same size which isolate each of our candidate changes. We then compare all five in a set of transfer learning experiments that look at our ability to use each of these datasets to improve performance on the eight downstream language understanding tasks in the SuperGLUE (Wang et al., 2019b) benchmark.

All five of our datasets are consistent with the task definition that was used in MNLI, which is in turn based on the definition introduced by Dagan et al. (2006). In this task, each example consists of a pair of short texts, called the *premise* and the *hypothesis*. The model is asked to read both texts and make a three-way classification decision: Given

---

[*]Work done while visiting Google.

Figure 1: The annotation interface for BASE.

the premise, would a reasonable person infer that hypothesis must be true (*entailment*), infer that that it must be false (*contradiction*), or decide that there is not enough information to make either inference (*neutral*)? While it is certainly not clear that this framing is optimal for pretraining, we leave a more broad-based exploration of task definitions for future work.

Our **BASE** data collection protocol (Figure 1) follows MNLI closely in asking annotators to read a premise sentence and then write three corresponding hypothesis sentences in empty text boxes corresponding to the three different labels (*entailment*, *contradiction*, and *neutral*). When an annotator follows this protocol, they produce three sentence pairs at once, all sharing a single premise.

Our **PARAGRAPH** protocol tests the effect of supplying annotators with complete paragraphs, rather than sentences, as premises. Longer texts offer the potential for discourse-level inferences, the addition of which should yield a dataset which is more difficult, more diverse, and less likely to contain trivial artifacts. However, reading full paragraphs could increase the time required to annotate a single example; time which could potentially be better spent constructing more sentence-level examples.

Our **EDITPREMISE** and **EDITOTHER** protocols test the effect of pre-filling a single seed text in each of the three text boxes that annotators are asked to fill out. By reducing the raw amount of typing required, this could allow annotators to produce good examples more quickly. By encouraging them to keep the three sentences similar, it could also indirectly facilitate the construction of minimal-pair-like examples that minimize artifacts, in the style of Kaushik et al. (2020). We test two variants of this idea: One uses a copy of the premise sentence as a seed text and the second retrieves a new

sentence from an existing corpus that is similar to the premise sentence, and uses that.

Our **CONTRAST** protocol tests the effect of adding artificial constraints on the kinds of hypothesis sentences annotators can write. Giving annotators difficult and varying constraints could encourage creativity and prevent annotators from falling into patterns in their writing that lead to easier or more repetitive data. However, as with the use of longer contexts in PARAGRAPH, this protocol risks substantially slowing the annotation process. We experiment with a procedure inspired by that used to create the language-and-vision dataset NLVR2 (Suhr et al., 2019), in which in which annotators must write sentences that are show some specified relationship (entailment or contradiction) to a given premise, but do not show that relationship to a second, similar, distractor premise.

In evaluations on transfer learning with the SuperGLUE benchmark, our BASE dataset and the datasets collected under our four new protocols offer substantial improvements in transfer ability over a plain RoBERTa or XLNet model, comparable to the gains seen with an equally-sized sample of MNLI. However, BASE reliably shows the strongest transfer results. This finding, combined with a low variance across runs, strongly suggests that none of these four interventions improves the suitability of NLI data for transfer learning. While this is a negative result for our primary focus on transfer, we also observe that all four of these methods are able to produce data of comparable subjective quality while significantly reducing the incidence of previously reported annotation artifacts, and that PARAGRAPH, EDITPREMISE, and EDITOTHER all accomplish this without significantly increasing the time cost of annotation.

2   Related Work

The observation that NLI data can be effective in pretraining was first reported for SNLI (Bowman et al., 2015) and MNLI by Conneau et al. (2017) on models pretrained from scratch on NLI data. This finding was replicated in the setting of multi-task pretraining by Subramanian et al. (2018). This was later extended to the context of *intermediate training*—where a model is pretrained on unlabeled data, then on relatively abundant labeled data (MNLI), and finally scarce task specific labeled data—by Phang et al. (2018), Clark et al. (2019), Liu et al. (2019a), Yang et al. (2019), and Liu et al.

MNLI (Training Set)
P: *Conceptually cream skimming has two basic dimensions - product and geography.* **H:** *Product and geography are what make cream skimming work.* **neutral**

BASE
P: *The board had also expressed concerns about the amounts of cash kept by SNC's Libyan office, at that time approximately $10 million, according to the company's chief financial officer.* **H:** *According to the board, the Libyan office should be holding more cash on hand.* **contradiction**

PARAGRAPH
P: *The paper, along with the "Washington Blade", was acquired by Window Media, LLC in 2001, and both were then sold to HX Media in 2007. Kat Long succeeded Trenton Straube as editor-in-chief in February 2009. The paper ceased publication in July 2009.* **H:** *Kal Long succeeded Trenton Straube as editor-in-chief in March 2019.* **contradiction**

EDITPREMISE
P: *This standpoint is believed to promote Deaf people's right to collective space within society to pass on their language and culture to future generations.* **H:** *This standpoint is believed to demote Deaf people's right to collective space within society.* **contradiction**

EDITOTHER
P: *Shobhona Sharma (born 5 February 1953) is a professor specializing in immunology, molecular biology, and biochemistry at the Tata Institute of Fundamental Research, Mumbai.* **H:** *Shobhona Sharma is also professor of mathematics at the Tata Institute of Fundamental Research, Mumbai.* **neutral**

CONTRAST
P: *Bengt Erik Johan Renvall (September 22, 1959 – August 24, 2015) was a Swedish dancer and choreographer active in the United States from 1978.* **H:** *He was a dancer in America in the 1970s.* **entailment**

Table 1: Randomly selected examples from the datasets under study. Neither the MNLI training set nor any of our collected data are filtered for quality in any way, and errors or debatable judgments are common in both.

(2019b) across a range of large pretrained models models and target language-understanding tasks. Similar results have been observed with transfer from the SocialIQA corpus (Sap et al., 2019) to target tasks centered on common sense. A small body of work including Mou et al. (2016), Bingel and Søgaard (2017) and Wang et al. (2019a) has explored the empirical landscape of which super-vised NLP tasks can offer effective pretraining for other supervised NLP tasks.

Existing NLI datasets have been built using a wide range of strategies: FraCaS (Cooper et al., 1996) and several targeted evaluation sets were constructed manually by experts from scratch. The RTE challenge corpora (Dagan et al., 2006, et seq.) primarily used expert annotations on top of existing premise sentences. SICK (Marelli et al., 2014) was created using a structured pipeline centered on asking crowdworkers to edit sentences in prescribed ways. MPE (Lai et al., 2017) uses a similar strategy, but constructs unordered *sets* of sentences for use as premises. SNLI (Bowman et al., 2015) introduced the method, used in MNLI, of asking crowdworkers to compose labeled hypotheses for a given premise. SciTail (Khot et al., 2018) and SWAG (Zellers et al., 2018) used domain-specific resources to pair existing sentences as potential entailment pairs, with SWAG additionally using trained models to identify examples worth annotating. There has been little work directly evaluating and comparing these many methods. In that absence, we focus on the SNLI/MNLI approach, because it has been shown to be effective for the collection of pretraining data and because its reliance on only crowdworkers and unstructured source text makes it simple to scale.

Two recent papers have investigated other methods that could augment the base MNLI protocol we study here. ANLI (Nie et al., 2019) collects new examples following this protocol, but adds an incentive for crowdworkers to produce sentence pairs on which a baseline system will perform poorly. Kaushik et al. (2020) introduce a method for expanding an already-collected dataset by making small edits to existing examples that change their labels, with the intent to produce minimally-different *minimal pairs* with differing labels. Both of these papers offer methodological changes that are potentially complementary to the changes we investigate here, and neither evaluates the impact of their methods on transfer learning. Since ANLI is large and roughly comparable with MNLI, we include it in our transfer evaluations here.

In addition to NLVR2, WinoGrande (Sakaguchi et al., 2019) also showed promising results from the use of artificial constraints during the annotation process for a different style of language-understanding dataset.

3   Data Collection

The basic interface for our tasks is similar to that used for SNLI and MNLI: We provide a premise from a preexisting text source and ask human annotators to provide three hypothesis sentences: one that says something true about the fact or situation in the prompt (*entailment*), one that says something that may or may not be true about the fact or situation in the prompt (*neutral*), and one that definitely does not say something true about the fact or situation in the prompt (*contradiction*).

BASE   In this baseline, modeled closely on the protocol used for MNLI, we show annotators a premise sentence and ask them to provide compose one new sentence for each label.

PARAGRAPH   Here, we use the same instructions as BASE, but with full paragraphs, rather than single sentences, as the supplied premises.

EDITPREMISE   Here, we pre-fill the three text boxes with editable copies of the premise sentence, and ask annotators to edit each text field to compose sentences that match the three different labels. Annotators are permitted to delete the pre-filled text.

EDITOTHER   Here, we follow the same procedure as EDITPREMISE, but rather than pre-filling the premise as a seed sentence, we instead use a similarity search method to retrieve a different sentence from the same source corpus that is similar to the premise.

CONTRAST   Here, we again retrieve a second sentence that is similar to the premise, but we display it as a *second premise* rather than using it to seed an editable text box. We then ask annotators to compose two new sentences: One sentence must be true *only* about the fact or situation in the first premise (that is, contradictory or neutral with respect to the second premise). The other sentence must be *false* only about the fact or situation in the first premise (and *true* or neutral with respect to the second premise). This yields an entailment pair and a contradiction pair, both of which use only the first premises, with the second premise serving only as a constraint on the annotation process. We could not find a sufficiently intuitive way to collect neutral sentence pairs under this protocol, and opted to use only two classes rather than increase the difficulty of an already unintuitive task.

3.1   Text Source

MNLI uses the small but stylistically diverse OpenANC corpus (Ide and Suderman, 2006) as its source for premise sentences, but uses nearly every available sentence from its non-technical sections. To avoid re-using premises, we instead draw on English Wikipedia.[1]

Similarity Search   The EDITOTHER and CONTRAST protocols require pairs of similar sentences as their inputs. To construct these, we assemble a heuristic sentence-matching system intended to generate pairs of highly similar sentences that can be minimally edited to construct entailments or contradictions: Given a premise, we retrieve its closest 10k nearest neighbors according to dot-product similarity over Universal Sentence Encoder (Cer et al., 2018) embeddings. Using a parser and an NER system, we then select those neighbors which share a subject noun phrase in common with the premise (dropping premises for which no such neighbors exist). From those filtered neighbors, we retrieve the single non-identical neighbor that has the highest overlap with the premise in both raw tokens and entity mentions, preferring sentences with similar length to the hypothesis.

3.2   The Annotation Process

We start data collection for each protocol with a pilot of 100 items, which are not included in the final datasets. We use these to refine task instructions and to provide feedback to our annotator pool on the intended task definition. We continue to provide regular feedback throughout the annotation process to clarify ambiguities in the protocols and to discourage the use of systematic patterns—such as consistently composing shorter hypotheses for entailments than for contradictions—that could make the resulting data artificially easy.

Annotators are allowed to skip prompts which they deem unusable for any reason. These generally involve either non-sentence strings that were mishandled by our sentence tokenizer or premises with inaccessible technical language. Skip rates ranged from about 2.5% for EDITOTHER to about 10% for CONTRAST (which can only be completed when the two premises are both comprehensible

---

[1]We use the 2019-06-20 downloadable version, extract the plain text with Apertium's WikiExtractor feature (Forcada et al., 2011), sentence-tokenize it with SpaCy (Honnibal and Montani, 2017), and randomly sample sentences (or paragraphs) for annotation.

	Label	Length $\mu\ (\sigma)$	Unique $\mu\ (\sigma)$
MNLI Gov. 8.5k			
premise	all labels	25.1 (13.4)	—
hypothesis	entailment	12.6 (5.1)	4.4 (5.1)
hypothesis	neutral	13.0 (5.3)	7.1 (5.3)
hypothesis	contradiction	12.0 (4.5)	5.7 (4.5)
BASE			
premise	all labels	23.3 (11.4)	—
hypothesis	entailment	10.6 (5.6)	2.3 (5.5)
hypothesis	neutral	10.5 (5.5)	4.5 (5.5)
hypothesis	contradiction	10.2 (5.1)	4.0 (5.1)
PARAGRAPH			
premise	all labels	66.7 (60.0)	—
hypothesis	entailment	13.0 (8.1)	2.3 (8.1)
hypothesis	neutral	12.9 (8.1)	4.1 (8.1)
hypothesis	contradiction	12.5 (7.9)	3.3 (7.9)
EDITPREMISE			
hypothesis	entailment	15.0 (8.9)	2.5 (8.9)
hypothesis	neutral	17.0 (9.8)	4.3 (9.8)
hypothesis	contradiction	15.3 (9.2)	3.3 (9.2)
EDITOTHER			
hypothesis	entailment	12.6 (6.3)	3.2 (6.3)
hypothesis	neutral	13.0 (6.8)	6.2 (6.8)
hypothesis	contradiction	12.7 (6.4)	4.7 (6.4)
CONTRAST			
hypothesis	entailment	7.9 (5.1)	2.5 (5.1)
hypothesis	contradiction	7.7 (4.9)	3.5 (4.9)

Table 2: Key text statistics. Premises are drawn from essentially the same distribution in all our tasks except PARAGRAPH, so are shown only once. The *Unique* column shows the number of tokens that appear in a hypothesis but not in the corresponding premise.

and sufficiently different from one another).

A pool of 19 professional annotators located in the United States worked on our tasks, with about ten working on each. As a consequence of this relatively small annotation team, many annotators worked under more than one protocol, which we ran consecutively. This introduces a modest potential bias against BASE, in that annotators start the later tasks having seen somewhat more feedback.

We do not use any kind of second-pass annotation process for quality control, and we neither designate a test set nor recommend these datasets for system evaluation. Our aim is to use our limited annotation time budget to collect the largest and best possible sample of (pre)training data, and we are motivated by work like Khetan et al. (2018) which calls into question the value of second-pass quality-control annotations for training data.

3.3 The Resulting Data

Using each protocol, we collect a training set of exactly 8,500 examples and a small validation set of at least 300 examples. Table 1 shows randomly chosen examples.[2]

Hypotheses are mostly fluent, full sentences that adhere to prescriptive writing conventions for US English. In constructing hypotheses, annotators often reuse words or phrases from the premise, but rearrange them, alter their inflectional forms, or substitute synonyms or antonyms. Hypotheses tend to differ from premises both grammatically and stylistically.

Table 2 shows some simple statistics on the collected text. Our clearest observation here is that the two methods that use seed sentences tend to yield longer hypotheses and tend not to show a clear relationship between hypothesis–premise token overlap and label. CONTRAST tends to produce shorter hypotheses.

Annotator Time Annotators completed each of the five protocols at a similar rate, taking 3–4 minutes per prompt. This goes against our expectations that the longer premises in PARAGRAPH should substantially slow the annotation process, and that the pre-filled text in EDITPREMISE and EDITOTHER should speed annotation. Since the relatively complex CONTRAST produces only two sentence pairs per prompt rather than three, it yields fewer examples per minute.

Label–Word Association Table 3 shows the three words in each dataset that are most strongly associated with specific labels, using the smoothed PMI method of Gururangan et al. (2018). We also include results for a baseline: an 8.5k-example sample from the *government documents* single-genre section of MNLI, which is meant to to be maximally comparable to the single-genre datasets we collect.

BASE shows similar associations to the original MNLI, but all four of our interventions reduce these associations at least slightly. The use of seed sentences, especially in EDITPREMISE, largely eliminates the strong association between negation and contradiction seen in MNLI, and no new strong associations appear to take its place.

---

[2]These datasets will be made public shortly, and a link will be made available in an updated version of this manuscript.

Word	Label	PMI	Counts
MNLI Gov. 8.5k			
never	contradiction	0.837	152/178
no	contradiction	0.828	342/426
any	contradiction	0.721	128/169
BASE			
never	contradiction	0.935	231/255
also	neutral	0.587	64/93
any	contradiction	0.585	46/64
PARAGRAPH			
never	contradiction	0.608	49/67
than	neutral	0.526	95/156
went	neutral	0.489	46/73
EDITPREMISE			
years	neutral	0.461	135/239
ago	neutral	0.443	17/21
eight	contradiction	0.437	13/15
EDITOTHER			
refused	contradiction	0.565	24/28
hardly	contradiction	0.507	16/17
later	neutral	0.482	48/77
MNLI Gov. 8.5k (two-class)			
no	contradiction	0.754	437/461
any	contradiction	0.689	193/208
only	contradiction	0.633	215/249
CONTRAST (two-class)			
only	contradiction	0.677	176/228
never	contradiction	0.635	73/90
no	contradiction	0.616	102/135

Table 3: The top three words most associated with specific labels in each dataset, sorted by the PMI between the word and the label. The *counts* column shows how many of the instances of each word occur in hypotheses matching the specified label. We compare the two-class CONTRAST with a two-class version of MNLI Gov.

4 Modeling Experiments

Our experiments generally compare models trained on ten different datasets: Each of the five 8.5k-example training sets introduced in this paper; the full 393k-example MNLI training set; the full 1.1m-example ANLI training set (which combines the SNLI training set, the MNLI training set, and the supplemental ANLI training examples);[3] 8.5k-example samples from the MNLI training set and from the combined ANLI training set, meant to control for the size differences between these existing datasets and our baselines; and finally a 8.5k-example sample from the government section of

the MNLI training set, meant to control (as much as possible) for the difference between our single-genre Wikipedia datasets and MNLI's relatively diverse text.

Our models are trained starting from pretrained RoBERTa (*large* variant; Liu et al., 2019b) or XLNet (*large*, *cased*; Yang et al., 2019). RoBERTa represented the state of the art on most of our target tasks as of the launch of our experiments. XLNet is competitive with RoBERTa on most tasks, and offers a natural replication. It can be used to better compare models trained on our data with models trained on ANLI: ANLI was collected with a model-in-the-loop procedure using RoBERTa that makes it difficult to interpret RoBERTa results.

We run our expemients using the `jiant` toolkit (Wang et al., 2019d), which implements the Super-GLUE tasks, MNLI, and ANLI, and in turn uses `transformers` (Wolf et al., 2019), AllenNLP (Gardner et al., 2017), and PyTorch (Paszke et al., 2017). To make it possible to train these large models on single consumer GPUs, we use small-batch ($b = 4$) training and a maximum total sequence length of 128 word pieces.[4] We train for up to 2 epochs for the very large ReCoRD, 10 epochs for the very small CB, COPA, and WSC, and 4 epochs for the remaining tasks. Except where noted, all results reflect the median final performance from three random restarts of training.[5]

Direct NLI Evaluations As a preliminary sanity check, Table 4 shows the results of evaluating models trained in each of the settings described above on their own validation sets, on the MNLI validation set, and on the expert-constructed GLUE diagnostic set (Wang et al., 2019c). As NLI classifiers trained on CONTRAST cannot produce the *neutral* labels used in MNLI, we evaluate them separately, and compare them with two-class variants of the MNLI models.

Our BASE data yields a model that performs somewhat worse than a comparable MNLI Gov. 8.5k model, both on their respective validation sets and on the full MNLI validation set. This suggests, at least tentatively, that the new annotations are less reliable than those in MNLI. This is disconcerting, but does not interfere with our key comparisons.

---

[3]In these runs, we use only the original ANLI validation set for evaluation and early stopping.

[4]We cut this to a slightly lower number on a few individual runs as needed to avoid memory constraints. Note that this potentially limits the gains observable for PARAGRAPH, which has a longer mean premise length of 66.7 words.

[5]Scripts implementing these experiments are available at github.com/nyu-mll/jiant/tree/nli-data.

Training Data	Self	MNLI	GLUE Diag.
BASE	84.8	81.5	40.5
PARAGRAPH	78.3	78.2	31.7
EDITPREMISE	82.9	79.8	35.5
EDITOTHER	82.5	82.6	33.9
MNLI8.5k	87.5	87.5	44.6
MNLIGov8.5k	87.7	85.4	40.7
ANLI8.5k	35.7	85.6	39.8
MNLI	90.4	**90.4**	49.2
ANLI	61.5	90.1	49.7
MNLI (two-class)	94.0	**94.0**	–
MNLI8.5k (two-class)	92.4	92.4	–
CONTRAST	91.6	80.6	–

Table 4: NLI modeling experiments with RoBERTa, reporting results on the validation sets for MNLI and for the task used for training each model (Self), and the GLUE diagnostic set. We compare the two-class CONTRAST with a two-class version of MNLI.

Training Data	Self	MNLI
BASE	57.9	52.2
PARAGRAPH	48.3	47.0
EDITPREMISE	40.4	39.4
EDITOTHER	45.1	50.7
MNLI8.5k	56.8	56.8
MNLIGov8.5k	63.7	53.9
ANLI8.5k	34.3	54.4
MNLI	62.0	**62.0**
ANLI	53.2	61.6
MNLI (two-class)	72.6	**72.6**
MNLI8.5k (two-class)	62.4	**62.4**
CONTRAST	56.9	55.9

Table 5: Results from RoBERTa hypothesis-only NLI classifiers on the vaidation sets for MNLI and for the datasets used in training.

The main conclusion we draw from these results is that none of the first three interventions improve performance on the out-of-domain GLUE diagnostic set, suggesting that they do not help in the collection of data that is both difficult and consistent with the MNLI label definitions. We also observe that the newer ANLI data yields worse performance than MNLI on the out-of-domain evaluation data when we control for dataset size.

Hypothesis-Only Models To further investigate the degree to which our hypotheses contain artifacts that reveal their labels, Table 5 shows results with single-input versions of our models trained on hypothesis-only versions of the datasets under study and evaluated on the datasets' validation sets.

Our first three interventions, especially EDIT-PREMISE, show much lower hypothesis-only performance than BASE. This drop is much larger

than the drop seen in in our standard NLI experimenst in the *Self* column of Table 4. This indicates that these these results cannot be explained away as a consequence of the lower quality of the evaluation sets for these three new datasets. This adds further evidence, alongside our PMI results, that these interventions reduce the presence of such artifacts. While we do not have a direct baseline for the two-class CONTRAST in this experiment, comparisons with MNLI 8.5k are consistent with the encouraging results seen above.

Transfer Evaluations For our primary evaluation, we use the training sets from our datasets in STILTs-style intermediate training (Phang et al., 2018): We fine-tune a large pretrained model on our collected data using standard fine-tuning procedures, then fine-tune a copy of the resulting model again on each of the target evaluation datasets we use. We then measure the aggregate performance of the resulting models across those evaluation datasets.

We evaluate on the target tasks in the Super-GLUE benchmark (Wang et al., 2019b): BoolQ (Clark et al., 2019), MultiRC (Khashabi et al., 2018), ReCoRD (Zhang et al., 2018), Commitment-Bank (CB; De Marneffe et al., 2019), Choice of Plausible Alternatives (COPA; Roemmele et al., 2011), Recognizing Textual Entailment (RTE; Dagan et al., 2006; Bar Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), the Winograd Schema Challenge (WSC; Levesque et al., 2012), and WiC (Pilehvar and Camacho-Collados, 2019). These tasks were selected to be difficult for BERT but relatively easy for non-expert humans, and are meant to replace the largely-saturated GLUE benchmark (Wang et al., 2019c).

SuperGLUE does not include labeled test data, and does not allow for substantial ablation analyses on its test sets. Since we have no single final model whose performance we aim to show off, we do not evaluate on the test sets. We also neither use any auxiliary WSC-format data when training our WSC model (as in Kocijan et al., 2019) nor artificially modify the task format. As has been observed elsewhere, we do not generally reach above-chance performance on that task without these extra techniques.

Results are shown in Table 6. Our first observation is that our overall data collection pipeline worked well for our purposes: Our BASE data yields models that transfer substantially better than

Intermediate-Training Data	Avg. $\mu(\sigma)$	BoolQ Acc.	CB F1/Acc.	COPA Acc.	MultiRC F1$_a$/EM	ReCoRD F1/EM	RTE Acc.	WiC Acc.	WSC Acc.
RoBERTa (large)									
None	67.3 (1.2)	84.3	83.1/89.3	90.0	70.0/27.3	86.5/85.9	85.2	**71.9**	64.4
Base	**72.2** (0.1)	84.4	**97.4/96.4**	**94.0**	71.9/33.3	86.1/85.5	88.4	70.8	**76.9**
Paragraph	70.3 (0.1)	84.7	**97.4/96.4**	90.0	70.4/29.9	**86.7/86.0**	86.3	70.2	67.3
EditPremise	69.6 (0.6)	83.0	92.3/92.9	89.0	71.2/31.2	86.4/85.7	85.6	71.0	65.4
EditOther	70.3 (0.1)	84.2	91.8/94.6	91.0	70.7/31.3	86.2/85.6	87.4	71.5	68.3
Contrast	69.2 (0.0)	84.1	93.1/94.6	87.0	71.4/29.5	84.8/84.1	84.5	71.5	67.3
MNLI8.5k	71.0 (0.6)	84.7	96.1/94.6	92.0	71.7/32.3	86.4/85.7	87.4	74.0	68.3
MNLIGov8.5k	70.9 (0.5)	**84.8**	**97.4/96.4**	92.0	71.4/32.0	86.2/85.6	86.3	71.6	70.2
ANLI8.5k	70.5 (0.3)	84.7	96.1/94.6	89.0	71.6/31.8	85.7/85.0	85.9	**71.9**	70.2
MNLI	70.0 (0.0)	85.3	89.0/92.9	88.0	**72.2**/35.4	84.7/84.1	89.2	71.8	66.3
ANLI	70.4 (0.9)	85.4	92.4/92.9	90.0	72.0/33.5	85.5/84.8	**91.0**	71.8	66.3
XLNet (large cased)									
None	62.7 (1.3)	82.0	83.1/89.3	76.0	69.9/26.8	80.9/80.1	69.0	65.2	63.5
Base	67.7 (0.0)	83.1	90.5/92.9	**89.0**	70.5/28.2	78.2/77.4	85.9	68.7	64.4
Paragraph	67.3 (0.0)	82.5	**90.8/94.6**	85.0	69.8/28.1	79.4/78.6	83.8	69.7	64.4
EditPremise	67.0 (0.4)	82.8	82.8/91.1	83.0	69.8/28.6	79.3/78.5	85.2	**70.2**	**65.4**
EditOther	67.2 (0.1)	82.9	84.4/91.1	87.0	70.2/29.1	79.4/78.6	85.6	69.7	63.5
Contrast	66.3 (0.6)	83.0	82.5/89.3	83.0	69.8/28.3	80.2/79.5	85.9	68.2	58.7
MNLI8.5k	67.6 (0.1)	83.5	89.5/92.9	88.0	69.4/28.3	79.5/78.6	86.3	69.3	62.5
MNLIGov8.5k	67.5 (0.3)	82.5	89.5/**94.6**	85.0	70.0/28.1	79.8/79.0	87.4	68.7	62.5
ANLI8.5k	67.2 (0.3)	83.4	86.3/91.1	83.0	69.3/28.9	**81.2/80.4**	85.9	70.1	63.5
MNLI	67.7 (0.1)	**84.0**	85.5/91.1	**89.0**	**71.5/31.0**	79.1/78.3	87.7	68.5	63.5
ANLI	**68.1** (0.4)	83.7	82.8/91.1	86.0	71.3/30.0	80.1/79.3	**89.5**	69.6	66.3

Table 6: Model performance on the SuperGLUE validation and diagnostic sets. The Avg. column shows the overall SuperGLUE score—an average across the eight tasks, weighting each task equally—as a mean and standard deviation across three restarts.

the plain RoBERTa or XLNet baseline, and at least slightly better than 8.5k-example samples of MNLI, MNLI Government or ANLI.

However, all four of our interventions yield *worse* transfer performance than Base. The variances across runs are small, and this pattern is consistent across both RoBERTa and XLNet, and across most individual target tasks.

We believe that this is a genuine negative result: At least under the broad experimental setting outline here, we find that none of these four interventions is helpful for transfer learning. We chose to collect 8,500 example samples because of the prior observation that this approximate amount was sufficient to show clear results on transfer learning, and we reproduce that finding here: Both MNLI 8.5k and the Base dataset yield large improvements over plain RoBERTa or XLNet through transfer learning. If any of our interventions were to be helpful in general, we would expect them to be harmless or helpful in our regime relative to Base. This is not what we observe.

We believe that this is the first study to evaluate ANLI as a pretraining task in transfer learning, and we observe that the large combined ANLI training

set yields consistently better transfer than the original MNLI dataset. However, we observe (to our surprise) that this result reverses when we control for length, with an 8.5k-example sample of MNLI yielding consistently better performance than an equivalently small sample of ANLI.

We also note that our best overall result uses only 8.5k NLI training examples, suggesting either that this size is enough to maximize the gains available through NLI pretraining, or that issues of forgetting during intermediate training make larger intermediate training sets more sensitive to optimization choices.

Finally, we replicate the finding from (Phang et al., 2018) that intermediate-task training with NLI data substantially reduces the variance across restarts seen in target task tuning.

5 Conclusion

Our chief results on transfer learning are conclusively negative: In this setting, all four interventions yield substantially worse transfer performance than our base MNLI data collection protocol. However, we also observe promising signs that all four of our interventions help to reduce the preva-

lence of artifacts in the generated hypotheses that reveal the label. This suggests that these methods may be valuable in the collection of high-quality *evaluation* data, if combined with additional validation methods to ensure high human agreement with the collected labels.

The need and opportunity that motivated this work remains compelling: Human-annotated data like MNLI has already proven itself as a valuable tool in teaching machines general-purpose skils for language understanding, and discovering ways to more effectively build and use such data could further accelerate the field's already fast progress toward robust, general-purpose language understanding technologies.

Further work along this line of research could productively follow a number of directions: General work on incentive structures and task design for crowdsourcing could help to address questions about how to collect data that is both creative and consistently labeled. Machine learning methods work on transfer learning could help to better understand and exploit the effects that drive the successes we have seen with NLI data so far. Finally, there remains room for further empirical work investigating the kinds of task definitions and data collection protocols most likely to yield positive transfer.

Acknowledgments

We thank the annotators who spent time and effort on this project and the many members of the natural language processing community at Google who provided feedback.

References

Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Textual Analysis Conference (TAC)*.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*. Springer.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. To appear in *Proceedings of Sinn und Bedeutung 23*. Data can be found at `https://github.com/mcdm/CommitmentBank/`.

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 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Association for Computational Linguistics.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Available at spacy.io.

Nancy Ide and Keith Suderman. 2006. Integrating linguistic resources: The national corpus model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.

Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. 2018. Learning from noisy singly-labeled

data. In *International Conference on Learning Representations (ICLR)*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the Winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint 1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 479–489, Austin, Texas. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint 1910.14599*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv preprint 1811.01088*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4462–4472, Hong Kong, China. Association for Computational Linguistics.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019b. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3261–3275. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019c. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Jason Phang, Edouard Grave, Haokun Liu, Najoung Kim, Phu Mon Htut, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2019d. `jiant` 1.2: A software toolkit for research on general-purpose text understanding models.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint 1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le.

2019. XLNet: Generalized autoregressive pre-training for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint 1810.12885*.