SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

Alex Wang*
New York University

Yada Pruksachatkun*
New York University

Nikita Nangia*
New York University

Amanpreet Singh
Facebook AI Research

Julian Michael
University of Washington

Felix Hill
DeepMind

Omer Levy
Facebook AI Research

Samuel R. Bowman
New York University

Abstract

In the last year, new models and methods for pretraining and transfer learning have driven striking performance improvements across a range of language understanding tasks. The GLUE benchmark, introduced one year ago, offers a single-number metric that summarizes progress on a diverse set of such tasks, but performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research. This paper recaps lessons learned from the GLUE benchmark and presents SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, improved resources, and a new public leaderboard. SuperGLUE is available at `super.gluebenchmark.com`.

1 Introduction

The past year has seen a surge of progress across many natural language processing (NLP) tasks, led by pretrained models like ELMo (Peters et al., 2018), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2019). The common thread connecting each of these contributions is that they couple self-supervised learning from massive unlabelled text corpora with a recipe for effectively adapting the resulting model to target tasks. The tasks that have proven amenable to this general approach include question answering, sentiment analysis, textual entailment, and parsing, among many others (Devlin et al., 2019; Kitaev and Klein, 2018). Besides their striking gains in performance on many such tasks, both ELMo and BERT have been recognized with best paper awards at major conferences and widespread deployment in industry.

In this context, the GLUE benchmark (organized by some of the same authors as this work, short for General Language Understanding Evaluation; Wang et al., 2019) has become a prominent evaluation framework and leaderboard for research towards general-purpose language understanding technologies. GLUE is a collection of nine language understanding tasks built on existing public datasets, together with private test data, an evaluation server, a single-number target metric, and an accompanying expert-constructed diagnostic set. GLUE was designed to provide a general-purpose evaluation of language understanding that covers a range of training data volumes, task genres, and

*Equal contribution. Correspondence: `glue-benchmark-admin@googlegroups.com`

Unpublished manuscript.

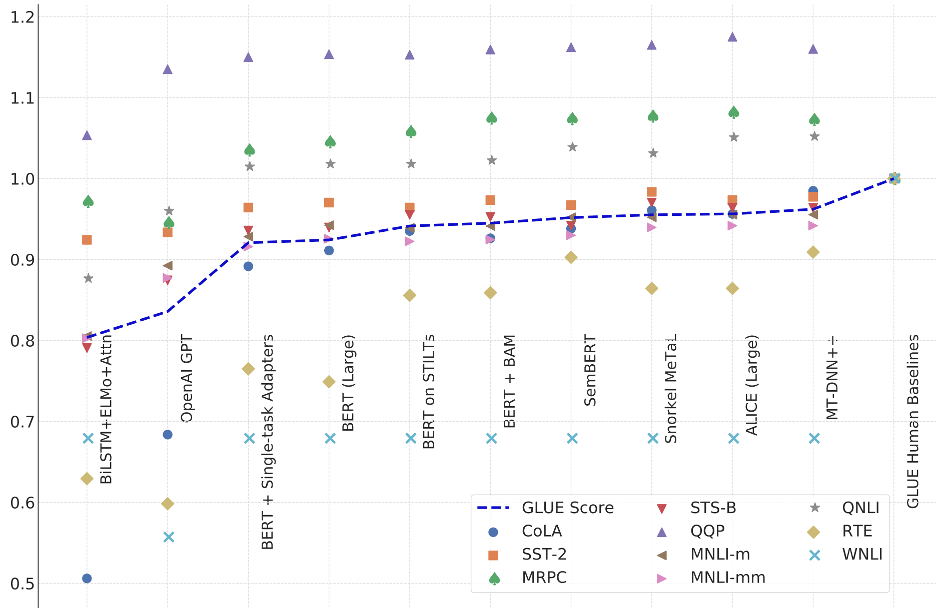

Figure 1: GLUE benchmark performance for submitted systems, rescaled to set human performance to 1.0, and shown both as a single number score and broken down into the nine constituent task performances. For tasks with multiple metrics, the result shown is based on an average of the metrics. More information on the tasks included in GLUE but not SuperGLUE can be found in Wang et al. (2019) and in Warstadt et al. (2018, CoLA), Socher et al. (2013, SST-2), Dolan and Brockett (2005, MRPC), Cer et al. (2017, STS-B), and Williams et al. (2018, MNLI), and Rajpurkar et al. (2016, the original data source for QNLI).

task formulations. We believe it was these aspects that made GLUE particularly appropriate for exhibiting the transfer-learning potential of approaches like OpenAI GPT and BERT.

With the progress seen over the last twelve months, headroom on the GLUE benchmark has shrunk dramatically, leaving GLUE somewhat limited in its ability to meaningfully quantify future improvements. While some tasks (Figure 1) and some linguistic phenomena (Figure 2) measured in GLUE remain difficult, the current state of the art GLUE Score (83.8 with the BERT-based MT-DNN system from Liu et al., 2019c) is only 3.3 points behind our estimate of global human performance (87.1 from Nangia and Bowman, 2019), and in fact exceeds this human performance estimate on three tasks.[2]

In response to this significant (and surprising) progress, this paper introduces an updated benchmark called SuperGLUE. SuperGLUE has the same fundamental objective as GLUE: to provide a simple hard-to-game benchmark for progress toward general-purpose language understanding technologies. We believe that SuperGLUE is significantly harder than GLUE, and is therefore more appropriate for measuring the impact of future developments in general-purpose models of language understanding.

What's New SuperGLUE follows the basic design of GLUE: It consists of a public leaderboard built around seven language understanding tasks, drawing on existing data, accompanied by a single-number performance metric and an analysis toolkit. It departs from GLUE in several ways:

- SuperGLUE retains only two of the nine GLUE tasks (one in a revised format), and replaces the remainder with a set of four new, more difficult tasks. These tasks were chosen to maximize difficulty and diversity, and were drawn from among those submitted to an open call for proposals.[3]

[2]The Quora Question Pairs, The Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005), and QNLI, an answer sentence selection task derived from SQuAD (Rajpurkar et al., 2016).

[3]`bit.ly/glue2cfp`

- Human performance estimates are included as part of the initial SuperGLUE benchmark release, and all of the included tasks have been selected to show a substantial headroom gap between a strong BERT-based baseline and human performance.
- The set of task formats (APIs) has expanded from sentence- and sentence-pair classification in GLUE to additionally include coreference resolution, sentence completion, and question answering in SuperGLUE.
- To facilitate the development of unified new methods for this more diverse set of tasks, SuperGLUE is distributed with a modular modeling toolkit for work on pretraining, multi-task learning, and transfer learning in NLP, built on PyTorch (Paszke et al., 2017) and AllenNLP (Gardner et al., 2017).
- The rules governing the SuperGLUE leaderboard differ from those governing GLUE in several ways, all meant to ensure fair competition, an informative leaderboard, and full credit assignment to data and task creators.

The SuperGLUE leaderboard and accompanying data and software tools are available now from `super.gluebenchmark.com` in a preliminary public trial version. Small changes to the benchmark may occur in response to late-breaking issues before the benchmark is frozen in a permanent state in early July 2019.

2 GLUE in Retrospect

Much work prior to GLUE has demonstrated that training neural models with large amounts of available supervision can produce representations that effectively transfer to a broad range of NLP tasks (Collobert and Weston, 2008; Dai and Le, 2015; Kiros et al., 2015; Hill et al., 2016; Conneau and Kiela, 2018; McCann et al., 2017; Peters et al., 2018). GLUE was presented as a formal challenge and leaderboard that could allow for straightforward comparison between such task-general transfer learning techniques. Other similarly-motivated benchmarks include SentEval (Conneau and Kiela, 2018), which evaluates fixed-size sentence embeddings, and DecaNLP (McCann et al., 2018), which recasts a set of target tasks into a general question-answering format and prohibits task-specific parameters. In contrast to these, GLUE provides a lightweight classification API and no restrictions on model architecture or parameter sharing, which seems to have been well suited to recent work in this area.

Since its release, GLUE has been used as a testbed and showcase by the developers of several influential models, including GPT (Radford et al., 2018) and BERT (Devlin et al., 2019). On GLUE, GPT and BERT achieved scores of 72.8 and 80.2 respectively, relative to 66.5 for an ELMo-based model (Peters et al., 2018) and 63.7 for the strongest baseline with no multitask learning or pretraining above the word level. These results demonstrate the value of sharing knowledge through self-supervised objectives that maximize the available training signal, modeling word occurrence conditioned on ever-richer context: from nearby unigrams (traditional distributional methods) to one-directional context (ELMo, GPT) and finally to bidirectional context (BERT).

Recently, Phang et al. (2018) showed that BERT could be improved by extending pretraining with labeled data related to a target task, and Liu et al. (2019c) showed further improvements using a specialized form of multi-task learning with parameter sharing. This results in an overall score of 83.8, an improvement of 3.3% over BERT, and a further sign of progress towards models with the expressivity and flexibility needed to acquire linguistic knowledge in one context or domain and apply it to others. Figure 1 shows progress on the benchmark to date.

However, limits to current approaches are also apparent in the GLUE suite. Top performance on Winograd-NLI (based on Levesque et al., 2012) is still at the majority baseline, with accuracy (65.1) far below human-level (95.9). Performance on the GLUE diagnostic entailment dataset, at 0.42 R_3, also falls far below the inter-annotator average of 0.80 R_3 reported in the original GLUE publication, with several categories of linguistic phenomena hard or adversarially difficult for top models (Figure 2). This suggests that even as unsupervised pretraining produces ever-better statistical summaries of text, it remains difficult to extract many details crucial to semantics without the right kind of supervision. Much recent work has observed this for NLI and QA (Jia and Liang, 2017; Naik et al., 2018; McCoy and Linzen, 2019; McCoy et al., 2019; Liu et al., 2019a,b).

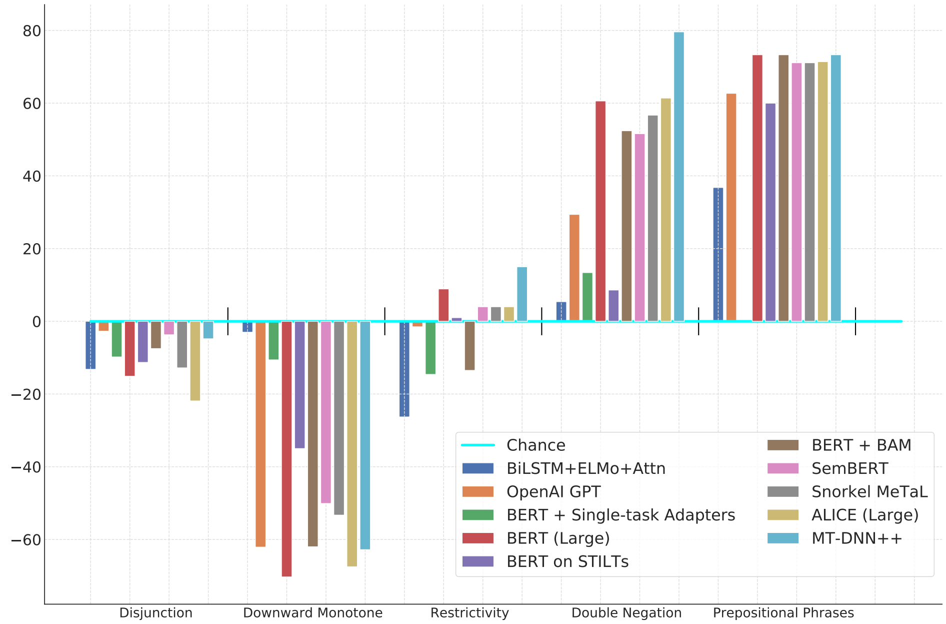

Figure 2: Performance of GLUE submissions on selected diagnostic categories, reported using the R_3 metric scaled up by 100, as in Wang et al. (2019, see paper for a description of the categories). While some initially difficult categories saw gains from advances on GLUE (e.g., double negation), others remained hard (restrictivity) or even adversarial (disjunction, downward monotone).

Table 1: The tasks included in SuperGLUE. *WSD* stands for word sense disambiguation, *NLI* is natural language inference, *coref.* is coreferece resolution, *SC* is sentence completion, and *QA* is question answering. For MultiRC, we list the number of total answers for 456/83/166 train/dev/test questions. The metrics for MultiRC are macro-average F1 over the set of questions (F1_m) and binary F1 on all answer-options (F1_a).

Corpus	**|Train|**	**|Dev|**	**|Test|**	**Task**	**Metrics**	**Text Sources**
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	SC	acc.	online blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$\text{F1}_m/\text{F1}_a$	various
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

Although models are fast approaching our (conservative) 87.3% estimate of non-expert human performance on GLUE—suggesting little remaining headroom on the benchmark—it seems unlikely that a machine capable of robust, human-level language understanding will emerge any time soon. To create a more challenging and stickier benchmark, we aim to focus SuperGLUE on datasets like Winograd-NLI: language tasks that are simple and intuitive for non-specialist humans but that pose a significant challenge to BERT and its friends.

3 Benchmark Tasks

Design Principles The goal of SuperGLUE is to provide a simple, robust evaluation of any method that can be uniformly applied to solve a broad range of language understanding tasks.To that end, we worked from these criteria when choosing tasks to include in the new benchmark:

- Task substance: Tasks should test a system's ability to understand English. We avoided prescribing a set of language understanding competencies and seeking out datasets to test those competencies. Instead, we opted to include any task that primarily involves language

Table 2: Selected development set examples from the tasks included in SuperGLUE. Text in **bold** represents part of the example format for each task. Text in *italics* is part of the model input. *Underlined* text is specially marked in the input. Text in `a monospaced font` is represents the expected model output.

CB

Text: *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?* **Hypothesis:** *they are setting a trend* **Entailment:** `Unknown`

COPA

Premise: *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?*
Alternative 1: *The sun was rising.* **Alternative 2:** *The grass was cut.* **Correct Alternative:** `1`

MultiRC

Paragraph: *(CNN) – Gabriel García Márquez, widely regarded as one of the most important contemporary Latin American authors, was admitted to a hospital in Mexico earlier this week, according to the Ministry of Health. The Nobel Prize recipient, known as "Gabo," had infections in his lungs and his urinary tract. He was suffering from dehydration, the ministry said. García Márquez, 87, is responding well to antibiotics, but his release date is still to be determined. "I wish him a speedy recovery." Mexican President Enrique Peña wrote on Twitter. García Márquez was born in the northern Colombian town of Aracataca, the inspiration for the fictional town of Macondo, the setting of the 1967 novel "One Hundred Years of Solitude." He won the Nobel Prize for literature in 1982 "for his novels and short stories, in which the fantastic and the realistic are combined in a richly composed world of imagination, reflecting a continent's life and conflicts," according to the Nobel Prize website. García Márquez has spent many years in Mexico and has a huge following there. Colombian President Juan Manuel Santos said his country is thinking of the author. "All of Colombia wishes a speedy recovery to the greatest of all time: Gabriel García Márquez," he tweeted. CNN en Español's Fidel Gutierrez contributed to this story.*
Question: *Whose speedy recover did Mexican President Enrique Peña wish on Twitter?*
Candidate answers: *Enrique Peña* (`F`), *Gabriel Garcia Marquez* (`T`), *Gabo* (`T`), *Gabriel Mata* (`F`), *Fidel Gutierrez* (`F`), *87* (`F`), *The Nobel Prize recipient* (`T`)

RTE

Text: *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*
Hypothesis: *Christopher Reeve had an accident.* **Entailment:** `False`

WSC

Text: *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** `False`

WiC

Context 1: *Room and board.* **Context 2:** *He nailed boards across the windows.*
Sense match: `False`

understanding to solve and meets the remaining criteria, trusting that diversity in task type, domain, etc. would naturally emerge.

- Task difficulty: Tasks should be beyond the scope of current state-of-the-art systems, but solvable by humans. We turned down tasks that required a significant amount of domain knowledge, e.g. reading medical notes, scientific papers, etc.
- Evaluability: Tasks must have an automatic performance metric that corresponds well to human judgment or performance. Many text generation tasks fail to meet this criteria due to issues surrounding automatic metrics like ROUGE and BLEU (Callison-Burch et al., 2006; Liu et al., 2016, i.a.).
- Public data: We required that tasks have existing public training data. We also preferred tasks for which we have access to or could create a test set with private labels.
- Task format: To avoid incentivizing the users of the benchmark to create complex task-specific model architectures, we preferred tasks that had relatively simple input and output formats. Previously we restricted GLUE to only include tasks involving single sentence or sentence pair inputs. With SuperGLUE, we expanded the scope to consider tasks with longer

inputs, leading to a set of tasks that requires understanding individual tokens in context, complete sentences, inter-sentence relations, and entire paragraphs.

- License: We required that task data be available under licences that allow use and redistribution for research purposes.

We disseminated a public call for proposals to the NLP community and received approximately 30 task submissions.[4] These proposals were then filtered according to the criteria above. Many proposals were not suitable due to licensing issues,[5] complex task formats,[6] and insufficient headroom. For each of the remaining tasks, we ran a simple BERT-based machine baseline and a human baseline, and filtered out tasks which were either too challenging for humans without extensive training[7] or too easy for our machine baselines.[8] The current version of SuperGLUE includes seven tasks, described in detail below and in summary in Tables 1 and 2.

CB The CommitmentBank (De Marneffe et al., 2019) is a corpus of short texts in which at least one sentence contains an embedded clause. Each of these embedded clauses is annotated with the degree to which we expect that the person who wrote the text is *committed* to the truth of the clause. The resulting task framed as three-class textual entailment on examples that are drawn from the Wall Street Journal, fiction from the British National Corpus, and Switchboard. Each example consists of a premise containing an embedded clause and the corresponding hypothesis is the extraction of that clause. We use a subset of the data that had inter-annotator agreement above 0.85. The data is imbalanced (relatively fewer *neutral* examples), so we evaluate using accuracy and F1, where for multi-class F1 we compute the unweighted average of the F1 per class.

COPA The Choice Of Plausible Alternatives (COPA, Roemmele et al., 2011) dataset is a causal reasoning task in which a system is given a premise sentence and two possible alternatives. The system must choose the alternative which has the more plausible causal relationship with the premise. The method used for the construction of the alternatives ensures that the task requires causal reasoning to solve. Examples either deal with alternative possible *causes* or alternative possible *effects* of the premise sentence, accompanied by a simple question disambiguating between the two instance types for the model. All examples are handcrafted and focus on topics from online blogs and a photography-related encyclopedia. Following the recommendation of the authors, we evaluate using accuracy.

MultiRC The Multi-Sentence Reading Comprehension dataset (MultiRC, Khashabi et al., 2018) is a true/false question-answering task. Each example consists of a context paragraph, a question about that paragraph, and a list of possible answers to that question which must be labeled as true or false. Question-answering (QA) is a popular problem with many datasets. We use MultiRC because of a number of desirable properties: (i) each question can have multiple possible correct answers, so each question-answer pair must be evaluated independent of other pairs, (ii) the questions are designed such that answering each question requires drawing facts from multiple context sentences, and (iii) the question-answer pair format more closely matches the API of other SuperGLUE tasks than span-based extractive QA does. The paragraphs are drawn from seven domains including news, fiction, and historical text. The evaluation metrics are F1 over all answer-options ($\mathrm{F1}_a$) and exact match of each question's set of answers (EM).

[4]We provide some examples of tasks that we considered but ultimately excluded in footnotes. We report on these excluded tasks only with the permission of their authors.

[5]Many medical text datasets are only accessible with explicit permission and credentials obtained from the creator.

[6]Tasks like QuAC (Choi et al., 2018a) and STREUSLE (Schneider and Smith, 2015) differed substantially from the format of other tasks in SuperGLUE, which we worried would incentivize users to spend significant effort on task-specific model designs, rather than focusing on general-purpose techniques.

[7]It was challenging to train annotators to do well on Quora Insincere Questions (`https://www.kaggle.com/c/quora-insincere-questions-classification/data`), Empathetic Reactions (Buechel et al., 2018), and a recast version of Ultra-Fine Entity Typing (Choi et al., 2018b, see Appendix A for details), leading to low human performance.

[8]BERT achieved very high or superhuman performance on Query Well-Formedness (Faruqui and Das, 2018) and PAWS (Zhang et al., 2019), Discovering Ongoing Conversations (Zanzotto and Ferrone, 2017), and GAP (Webster et al., 2018).

RTE The Recognizing Textual Entailment (RTE) datasets come from a series of annual competitions on textual entailment, the problem of predicting whether a given premise sentence entails a given hypothesis sentence (also known as natural language inference, NLI). RTE was previously included in GLUE, and we use the same data and format as before: We merge data from RTE1 (Dagan et al., 2006), RTE2 (Bar Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009).[9] All datasets are combined and converted to two-class classification: *entailment* and *not_entailment*. Of all the GLUE tasks, RTE was among those that benefited from transfer learning the most, jumping from near random-chance performance (∼56%) at the time of GLUE's launch to 85% accuracy (Liu et al., 2019c) at the time of writing. Given the eight point gap with respect to human performance, however, the task is not yet solved by machines, and we expect the remaining gap to be difficult to close.

WiC The Word-in-Context (WiC, Pilehvar and Camacho-Collados, 2019) dataset supports a word sense disambiguation task cast as binary classification over sentence pairs. Given two sentences and a polysemous (sense-ambiguous) word that appears in both sentences, the task is to determine whether the word is used with the same sense in both sentences. Sentences are drawn from WordNet (Miller, 1995), VerbNet (Schuler, 2005), and Wiktionary. We follow the original work and evaluate using accuracy.

WSC The Winograd Schema Challenge (WSC, Levesque et al., 2012) is a reading comprehension task in which a system must read a sentence with a pronoun and select the referent of that pronoun from a list of choices. Given the difficulty of this task and the headroom still left, we have included WSC in SuperGLUE and recast the dataset into its coreference form. The task is cast as a binary classification problem, as opposed to N-multiple choice, in order to isolate the model's ability to understand the coreference links within a sentence as opposed to various other strategies that may come into play in multiple choice conditions. With that in mind, we create a split with 65% negative majority class in the validation set, reflecting the distribution of the hidden test set, and 52% negative class in the training set. The training and validation examples are drawn from the original Winograd Schema dataset (Levesque et al., 2012), as well as those distributed by the affiliated organization *Commonsense Reasoning*.[10] The test examples are derived from fiction books and have been shared with us by the authors of the original dataset. Previously, a version of WSC recast as NLI as included in GLUE, known as WNLI. No substantial progress was made on WNLI, with many submissions opting to submit only majority class predictions. WNLI was made especially difficult due to an adversarial train/dev split: Premise sentences that appeared in the training set sometimes appeared in the development set with a different hypothesis and a flipped label. If a system memorized the training set without meaningfully generalizing, which was easy due to the small size of the training set, it could perform far *below* chance on the development set. We remove this adversarial design in the SuperGLUE version of WSC by ensuring that no sentences are shared between the training, validation, and test sets.

However, the validation and test sets come from different domains, with the validation set consisting of ambiguous examples such that changing one non-noun phrase word will change the coreference dependencies in the sentence. The test set consists only of more straightforward examples, with a high number of noun phrases (and thus more choices for the model), but low to no ambiguity.

The SuperGLUE Score As with GLUE, we seek to give a sense of aggregate system performance over all tasks by introducing the SuperGLUE score: an average of all task scores. We do not weight data-rich tasks more heavily than data-poor tasks to avoid concentrating research efforts on data-rich tasks, for which existing methods already perform relatively well. For Commitment Bank and MultiRC, we first take the average of the task's metrics, e.g. for MultiRC we first average $\mathrm{F1}_a$ and EM before using the resulting number as a single term in the overall average.

Diagnostics In addition to the task test sets, GLUE provides an expert-constructed diagnostic test set for the automatic analysis of textual entailment output. Each entry in the diagnostic set is a sentence pair labeled with a three-way entailment relation— *entailment*, *neutral*, or *contradiction*, matching the MultiNLI (Williams et al., 2018) label set—and tagged with labels that indicate a broad set of linguistic phenomena that characterize the relationship between the two sentences. Submissions

[9]RTE4 is not publicly available, while RTE6 and RTE7 don't conform to the standard NLI task.

[10]`http://commonsensereasoning.org/disambiguation.html`

to the GLUE leaderboard were requested to include predictions from the submission's MultiNLI classifier on the diagnostic set, and analyses of the results were shown alongside the main leaderboard.

Since the diagnostic task remains difficult for top models, we retain it in SuperGLUE. However, since MultiNLI is not part of SuperGLUE, we collapse *contradiction* and *neutral* into a single *not_entailment* label, and request that submissions include predictions on the collapsed diagnostic set from their *RTE* model.

To validate the data, we also collect a fresh set of non-expert annotations estimate human performance on the diagnostic dataset. We follow the same procedure that was used for estimating human performance on all the SuperGLUE tasks (Section 5.2). We estimate an accuracy of 88% and a Mathhew's correlation coefficient (MCC, the two-class variant of the R_3 coefficient used in GLUE) of 0.77.

4 Using SuperGLUE

Software Tools To facilitate using SuperGLUE, we also release a toolkit,[11] built around PyTorch (Paszke et al., 2017) and components from AllenNLP (Gardner et al., 2017), which implements our baselines and supports the evaluation of custom models and training methods on the benchmark tasks. The toolkit includes include existing popular pretrained models such as OpenAI GPT and BERT and uses a modular design for fast experimenting with different model components as well as multitask training.

Eligibility Any system or method that can produce predictions for the tasks in SuperGLUE is eligible for submission, subject to the data-use and submission frequency policies stated immediately below. There are no restrictions on the type of methods that may be used, and there is no requirement that any form of parameter sharing or shared initialization be used across the tasks in the benchmark.

Data Data for the SuperGLUE tasks are available for download through the SuperGLUE site and through a download script included with the software toolkit. Each task comes with a standardized training set, development set, and *unlabeled* test set.

Submitted systems may use any public or private data when developing their systems, with a few exceptions: Systems may only use the SuperGLUE-distributed versions of the SuperGLUE task datasets, as these use different train/validation/test splits from other public versions in some cases. Systems also may not use the unlabeled test data for the SuperGLUE tasks in system development in any way, and may not build systems that share information across separate test examples in any way.

The Leaderboard To compete on SuperGLUE, authors must submit a zip file containing predictions from their system to the SuperGLUE website to be scored by an auto-grader. By default, all submissions are private. To submit a system to the public leaderboard, one must score it and fill out a short additional form supplying either a short description or a link to a paper. Anonymous submissions are allowed, but will only be posted only when they are accompanied by an (anonymized) full paper. Users are limited to a maximum of two submissions per day and six submissions per month.

Further, to ensure reasonable credit assignment since SuperGLUE builds very directly on prior work, we ask the authors of submitted systems to directly name and cite the specific datasets that they use, *including the SuperGLUE datasets*. We will enforce this as a requirement for papers listed on the leaderboard.

Table 3: Preliminary baseline performance on the SuperGLUE task test sets. For GAP we report F1, as well as the bias score. For CB we report accuracy and macro-average F1. For MultiRC we report F1 on all answer-options and exact match of each question's set of correct answers. All values are scaled by 100. The *Avg* column is the average performance on all SuperGLUE tasks. The bolded numbers reflect the best machine performance on task. * MultiRC has a staggered test set release, and these results evaluate on a subset of the data we use, and thus are not directly comparable.

Model	**Avg**	**CB**	**COPA**	**MultiRC**	**RTE**	**WiC**	**WSC**
Metrics		**Acc./F1**	**Acc.**	**F1$_a$/EM**	**Acc.**	**Acc.**	**Acc.**
Most Frequent Class	46.9	48.4/21.7	50.0	61.1 / 0.3	50.4	50.0	65.1
CBOW	48.6	69.2/47.6	52.2	38.8 / 0.5	50.4	50.0	61.0
BERT	68.0	84.4/80.6	69.8	66.2 / 22.2	73.2	**70.4**	**67.8**
BERT++	**70.5**	**88.4/82.7**	77.4	66.2 / 22.2	77.6	**70.4**	**67.8**
Outside Best	-	- / -	**84.4**	**70.4*/24.5***	**82.7**	-	-
Human (estimate)	89.6	98.9/95.8	100.0	81.8*/51.9*	93.6	80.0	100.0

5 Baselines

5.1 Model

Our main baselines are built around BERT, variants of which are the most successful approach on GLUE to date. Specifically, we use the BERT-LARGE-CASED variant.[12] Following standard practice from Devlin et al. (2019), for each task, we use the simplest possible architecture on top of BERT, described in brief below.

For classification tasks with sentence-pair inputs (WiC, RTE, CB), we concatenate the sentences with a [SEP] token, feed the fused input to BERT, and use a logistic regression classifier that sees the representation corresponding to [CLS]. For WiC only, we also concatenate the representation of the marked word to the [CLS] representation. For COPA and MultiRC, for each answer choice, we similarly concatenate the context with that answer choice and feed the resulting sequence into BERT to produce an answer representation. For COPA, we project these representations into a scalar, and take as the answer the choice with the highest associated scalar. For MultiRC, because each question can have more than one correct answer, we feed each answer representation into a logistics regression classifier. For WSC, which is a a span-based task, we use a model inspired by Tenney et al. (2019). Given the BERT representation for each word in the original sentence, we get span representations of the pronoun and noun phrase via a self-attention span-pooling operator (Lee et al., 2017), before feeding it into a logistic regression classifier.

For training, we use the procedure specified in Devlin et al. (2019). Specifically, we use Adam (Kingma and Ba, 2014) with an initial learning rate of 10^{-5} and fine-tune for a maximum of 10 epochs. We fine-tune a copy of the pretrained BERT model separately for each task, and leave the development of multi-task learning models to future work. The results for this model are shown in the BERT row in Table 3.

We also report results using BERT with additional training on related datasets before fine-tuning on the SuperGLUE tasks, following the STILTs two-stage style of transfer learning (Phang et al., 2018). Given the productive use of MultiNLI in pretraining and intermediate fine-tuning of pretrained language models (Conneau et al., 2017; Phang et al., 2018, i.a.), for CB and RTE, we use MultiNLI as a transfer task by first using the above procedure on MultiNLI. Similarly, given the similarity of COPA to SWAG (Zellers et al., 2018), we first fine-tune BERT on SWAG. These results are reported as BERT++. For all other tasks, we reuse the results of BERT fine-tuned on just that task.

We also include a baseline where for each task we simply predict the majority class, as well as a bag-of-words baseline where each input is represented as an average of its tokens' GloVe word vectors (300-dimensional and trained on 840B Common Crawl tokens, Pennington et al., 2014).

[11] `https://jiant.info/`

[12] We use the PyTorch implementation by HuggingFace: `https://github.com/huggingface/pytorch-pretrained-BERT`

Finally, we also list the best known result on each task to date. We omit these numbers for tasks which we recast (WSC), resplit (CB), or achieve the best known result (WiC). The outside results for COPA, MultiRC, and RTE are from Sap et al. (2019), Trivedi et al. (2019), and Liu et al. (2019c) respectively.

5.2 Human Performance

Several datasets have non-expert human performance baselines already available. Pilehvar and Camacho-Collados (2019) provide an estimate for human performance on WiC in their paper. Similarly, Khashabi et al. (2018) also provide a human performance estimate with the release of MultiRC. Nangia and Bowman (2019) establish human performance for RTE. For the remaining SuperGLUE datasets, including the diagnostic set, we establish an estimate for human performance by hiring crowdworker annotators through the Amazon's Mechanical Turk platform[13] to reannotate a sample of each test set.

We follow a two step procedure where a crowd worker completes a short training phase before proceeding to the annotations phase, modeled after the method used by Nangia and Bowman (2019) for GLUE. The training phase uses 30 examples taken from the development set of the task. During training, workers are provided with instructions on the task, they are linked to an FAQ page, and are asked to annotate examples from the development set. After answering each example, workers are also asked to check their work by clicking on a "Check Work" button which reveals the ground truth label.

After the training phase is complete, we provide the qualification to work on the annotation phase to all workers who annotated a minimum of five examples, i.e. completed five HITs during training and achieved performance at, or above the median performance across all workers during training. In the annotation phase, workers are provided with the same instructions as the training phase, and are linked to the same FAQ page. The instructions for all tasks are provided in Appendix A.

For the annotation phase we randomly sample 100 examples from the task's test set, with the exception of WSC where we annotate the full 146-example test set. For each example, we collect redundant annotations from five workers and take a majority vote to estimate human performance. For task-specific details on how we present the tasks to annotators and calculate human performance numbers, refer to Appendix B.

For both training and annotation phases across all tasks, the average pay rate is $22.55/hr.[14]

6 Discussion

The results for all baselines are shown in Table 3. As expected, we observe that our simple baselines of predicting the most frequent class and CBOW do not perform well overall, achieving near chance performance for several of the tasks. Using BERT increases the average SuperGLUE score by nearly 20 points. On CB, we achieve strong accuracy and F1 scores of 84.4 and 80.6 respectively. We get further gains by training on related tasks like MultiNLI and SWAG.

However, our best pretraining baselines still lag substantially behind human performance. On average, there is nearly 20 point gap between BERT++ and human performance on SuperGLUE. The largest of the gaps is on WSC, with a 32.2 point accuracy difference between the best model and human performance. The smallest margins are on CB, RTE, and WiC with respective gaps of 11.8, 10.9, and 9.6 points.

While overall there is headroom on SuperGLUE, this gap is not *massive*, even though our design principles for SuperGLUE aim to maximize difficulty and we only included the hardest tasks from among those submitted. We believe this a reflection of the fact that current state-of-the-art models, like BERT, are genuinely fairly effective at sentence understanding in non-adversarial settings.

Ultimately though, we do believe this gap will be challenging to close. On WSC and COPA, human performance is perfect. On three other tasks, it is in the mid-to-high 90s. Given the estimated

[13] `https://www.mturk.com/`

[14] This estimate is taken from `https://turkerview.com`, where crowd workers self-report their hourly income on tasks.

headroom, there is plenty of space to test new creative approaches on a broad suite of difficult NLP tasks with SuperGLUE.

7 Acknowledgments

We thank the original authors of the included datasets in SuperGLUE for their cooperation in the creation of the benchmark, and we are also grateful to those who proposed task and datasets that we ultimately could not include in SuperGLUE.

This work was made possible in part by a donation to NYU from Eric and Wendy Schmidt made by recommendation of the Schmidt Futures program. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan V GPU used at NYU for this research. AW is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1342536. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL recognising textual entailment challenge. 2006.

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

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL `https://www.aclweb.org/anthology/S17-2001`.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D18-1241`.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia, July 2018b. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P18-1009`.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, May 2018. European Language Resource Association. URL `https://www.aclweb.org/anthology/L18-1269`.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings* *of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL `https://www.aclweb.org/anthology/D17-1070`.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer, 2006.

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3079–3087. Curran Associates, Inc., 2015. URL `http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning.pdf`.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. 2019. 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. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of IWP*, 2005.

Manaal Faruqui and Dipanjan Das. Identifying well-formed natural language questions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 798–803, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D18-1091`.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. *arXiv preprint 1803.07640*, 2017.

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

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1162. URL `https://www.aclweb.org/anthology/N16-1162`.

Robin Jia and Percy Liang. 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, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL `https://www.aclweb.org/anthology/D17-1215`.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint 1412.6980*, 2014.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

Nikita Kitaev and Dan Klein. Multilingual constituency parsing with self-attention and pre-training. *arXiv preprint 1812.11760*, 2018.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL `https://www.aclweb.org/anthology/D17-1018`.

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

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1230. URL `https://www.aclweb.org/anthology/D16-1230`.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019a.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019b.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint 1901.11504*, 2019c.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

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

Richard T. McCoy and Tal Linzen. Non-entailed subsequences as a challenge for natural language inference. In *Proceedings of the Society for Computational in Linguistics (SCiL) 2019*, 2019. URL `https://scholarworks.umass.edu/scil/vol2/iss1/46/`.

George A Miller. WordNet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Aakanksha Naik, Abhilasha Ravichander, Norman M. Sadeh, Carolyn Penstein Rosé, and Graham Neubig. Stress test evaluation for natural language inference. In *COLING*, 2018.

Nikita Nangia and Samuel R. Bowman. A conservative human baseline estimate for GLUE: People still (mostly) beat machines. Unpublished manuscript available at gluebenchmark.com, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://www.aclweb.org/anthology/D14-1162`.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://www.aclweb.org/anthology/N18-1202`.

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

Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. Unpublished ms. available through a link at `https://blog.openai.com/language-unsupervised/`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1264. URL `http://aclweb.org/anthology/D16-1264`.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019.

Nathan Schneider and Noah A Smith. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, 2015.

Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, Philadelphia, PA, USA, 2005. AAI3179808.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642, 2013.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. 2019. URL `https://openreview.net/forum?id=SJzSgnRcKX`.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. Repurposing entailment for multi-hop question answering tasks, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJ4km2R5t7`.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint 1805.12471*, 2018.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/N18-1101`.

Fabio Massimo Zanzotto and Lorenzo Ferrone. Have you lost the thread? discovering ongoing conversations in scattered dialog blocks. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(2):9, 2017.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. pages 93–104, October-November 2018. URL `https://www.aclweb.org/anthology/D18-1009`.

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

A Instructions to Crowd Workers

A.1 Training Phase Instructions

We provide workers with instructions about the training phase. An example of these instructions is given Table 4. These training instructions are the same across tasks, only the task name in the instructions is changed.

A.2 Task Instructions

During training and annotation for each task, we provide workers with brief instructions tailored to the task. We also link workers to an FAQ page for the task. Tables 5, 6, and 7 show the instructions we used for all four tasks: COPA, CommitmentBank, and WSC respectively. The instructions given to crowd workers for annotations on the diagnostic dataset are shown in Table 8.

We collected data to produce conservative estimates for human performance on several tasks that we did not ultimately include in SuperGLUE, including GAP (Webster et al., 2018), PAWS (Zhang et al., 2019), Quora Insincere Questions,[15] Ultrafine Entity Typing (Choi et al., 2018b), and Empathetic Reactions datasets (Buechel et al., 2018). The instructions we used for these tasks are shown in Tables 9, 10, 11, 12, and 13.

Ultrafine Entity Typing We cast the task into a binary classification problem to make it an easier task for non-expert crowd workers. We work in cooperation with the authors of the dataset (Choi et al., 2018b) to do this reformulation: We give workers one possible tag for a word or phrase and asked them to classify the tag as being applicable or not.

The authors used WordNet (Miller, 1995) to expand the set of labels to include synonyms and hypernyms from WordNet. They then asked five annotators to validate these tags. The tags from this validation had high agreement, and were included in the publicly available Ultrafine Entity Typing dataset,[16] This constitutes our set of positive examples. The rest of the tags from the validation procedure that are not in the public dataset constitute our negative examples.

GAP For the Gendered Ambiguous Pronoun Coreference task (GAP, Webster et al., 2018), we simplified the task by providing noun phrase spans as part of the input, thus reducing the original structure prediction task to a classification task. This task was presented to crowd workers as a three way classification problem: Choose span A, B, or neither.

B Human Performance Baseline on SuperGLUE

For WSC and COPA we provide annotators with a two way classification problem. We then use majority vote across annotations to calculate human performance.

CommitmentBank We follow the authors in providing annotators with a 7-way classification problem. We then collapse the annotations into 3 classes by using the same ranges for bucketing used

[15] `https://www.kaggle.com/c/quora-insincere-questions-classification/data`

[16] `https://homes.cs.washington.edu/~eunsol/open_entity.html`

by De Marneffe et al. (2019). We then use majority vote to get human performance numbers on the task.

Furthermore, for training on CommitmentBank we randomly sample examples from the low inter-annotator agreement portion of the CommitmentBank data that is not included in the SuperGLUE version of the task. These low agreement examples are generally harder to classify since they are more ambiguous.

Diagnostic Dataset Since the diagnostic dataset does not come with accompanying training data, we train our workers on examples from RTE's development set. RTE is also a textual entailment task and is the most closely related task in SuperGLUE. Providing the crowd workers with training on RTE enables them to learn label definitions which should generalize to the diagnostic dataset.

C Results on Excluded Tasks

During the process of selecting tasks for SuperGLUE, we collect human performance baselines and run BERT-based machine baselines for some tasks that we ultimately exclude from our task list. We choose to exclude these tasks because our BERT baseline performs better than our human performance baseline or if the gap between human and machine performance is small.

On Quora Insincere Questions[17] our BERT baseline outperforms our human baseline by a small margin: an F1 score of 67.2 versus 66.7 for BERT and human baselines respectively. Similarly, on the Empathetic Reactions dataset (Buechel et al., 2018), BERT outperforms our human baseline, where BERT's predictions have a Pearson correlation of 0.45 on empathy and 0.55 on distress, compared to 0.45 and 0.35 for our human baseline. For PAWS-Wiki, Zhang et al. (2019) report that BERT achieves an accuracy of 91.9%, while our human baseline achieved 84% accuracy. These three tasks are excluded from SuperGLUE since our, admittedly conservative, human baselines are worse than machine performance. Our human performance baselines are subject to the clarity of our instructions (all instructions can be found in Appendix A), and crowd workers engagement and ability.

For the Query Well-Formedness (Faruqui and Das, 2018) task, the authors set an estimate human performance at 88.4% accuracy. Our BERT baseline model reaches an accuracy of 82.3%. While there is a positive gap on this task, the gap was smaller than we were were willing to tolerate. Similarly, on our recast version of the Ultrafine Entity Typing (Choi et al., 2018b), we observe too small a gap between human (60.2 F1) and machine performance (55.0 F1). Our recasting for this task is described in Appendix A.2. On GAP (Webster et al., 2018), when taken as a classification problem without the related task of span selection (details in A.2), BERT performs (91.0 F1) comparably to our human baseline (94.9 F1). Given this small margin, we also exclude GAP from SuperGLUE.

On Discovering Ongoing Conversations (Zanzotto and Ferrone, 2017), our BERT baseline achieves an F1 of 51.9 on a version of the task cast as sentence pair classification (given two snippets of texts from plays, determine if the second snippet is a continuation of the first). This dataset is very class imbalanced (90% negative), so we also experimented with a class-balanced version on which our BERT baselines achieves 88.4 F1. Qualitatively, we also found the task challenging for humans as there was little context for the text snippets and the examples were drawn from plays using early English. Given this fairly high machine performance and challenging nature for humans, we exclude this task from SuperGLUE.

Instructions tables begin on the following page.

[17] `https://www.kaggle.com/c/quora-insincere-questions-classification/data`

Table 4: The instructions given to crowd-sourced worker describing the training phase for the Choice of Plausible Answers (COPA) task.

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

This project is a **training task** that needs to be completed before working on the main project on AMT named Human Performance: Plausible Answer. Once you are done with the training, please proceed to the main task! The qualification approval is not immediate but we will add you to our qualified workers list within a day.

In this training, you must answer the question on the page and then, to see how you did, click the **Check Work** button at the bottom of the page before hitting Submit. The Check Work button will reveal the true label. Please use this training and the provided answers to build an understanding of what the answers to these questions look like (the main project, Human Performance: Plausible Answer, does not have the answers on the page).

Table 5: Task-specific instructions for Choice of Plausible Alternatives (COPA). These instructions were provided during both training and annotation phases.

Plausible Answer Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with a prompt sentence and a question. The question will either be about what caused the situation described in the prompt, or what a possible effect of that situation is. We will also give you two possible answers to this question. Your job is to decide, given the situation described in the prompt, which of the two options is a more plausible answer to the question:

In the following example, option 1. is a more plausible answer to the question about what caused the situation described in the prompt,

> *The girl received a trophy.*
> *What's the CAUSE for this?*
> 1. *She won a spelling bee.*
> 2. *She made a new friend.*

In the following example, option 2. is a more plausible answer the question about what happened because of the situation described in the prompt,

> *The police aimed their weapons at the fugitive.*
> *What happened as a RESULT?*
> 1. *The fugitive fell to the ground.*
> 2. *The fugitive dropped his gun.*

If you have any more questions, please refer to our FAQ page.

Table 6: Task-specific instructions for Commitment Bank. These instructions were provided during both training and annotation phases.

Speaker Commitment Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with a prompt taken from a piece of dialogue, this could be a single sentence, a few sentences, or a short exchange between people. Your job is to figure out, based on this first prompt (on top), how certain the speaker is about the truthfulness of the second prompt (on the bottom). You can choose from a 7 point scale ranging from (1) completely certain that the second prompt is true to (7) completely certain that the second prompt is false. Here are examples for a few of the labels:

Choose 1 (certain that it is true) if the speaker from the first prompt definitely believes or knows that the second prompt is true. For example,

> *"What fun to hear Artemis laugh. She's such a serious child. I didn't know she had a sense of humor."*
>
> *"Artemis had a sense of humor"*

Choose 4 (not certain if it is true or false) if the speaker from the first prompt is uncertain if the second prompt is true or false. For example,

> *"Tess is committed to track. She's always trained with all her heart and soul. One can only hope that she has recovered from the flu and will cross the finish line."*
>
> *"Tess crossed the finish line."*

Choose 7 (certain that it is false) if the speaker from the first prompt definitely believes or knows that the second prompt is false. For example,

> *"Did you hear about Olivia's chemistry test? She studied really hard. But even after putting in all that time and energy, she didn't manage to pass the test".*
>
> *"Olivia passed the test."*

If you have any more questions, please refer to our FAQ page.

Table 7: Task-specific instructions for Winograd Schema Challenge (WSC). These instructions were provided during both training and annotation phases.

Winograd Schema Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with a sentence that someone wrote, with one bolded pronoun. We will then ask if you if the pronoun refers to a specific word or phrase in the sentence. Your job is to figure out, based on the sentence, if the bolded pronoun refers to this selected word or phrase:

Choose Yes if the pronoun refers to the selected word or phrase. For example,

> *"I put the cake away in the refrigerator. It has a lot of butter in it."*
> *Does **It** in "It has a lot" refer to **cake**?*

Choose No if the pronoun does not refer to the selected word or phrase. For example,

> *"The large ball crashed right through the table because it was made of styrofoam."*
> *Does **it** in "it was made" refer to **ball**?*

If you have any more questions, please refer to our FAQ page.

Table 8: Task-specific instructions for the diagnostic dataset. These instructions were provided during both training and annotation phases.

Textual Entailment Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with a prompt taken from an article someone wrote. Your job is to figure out, based on this correct prompt (the first prompt, on top), if another prompt (the second prompt, on bottom) is also necessarily true:

Choose True if the event or situation described by the first prompt definitely implies that the second prompt, on bottom, must also be true. For example,

- *"Murphy recently decided to move to London."*
  "Murphy recently decided to move to England."
  (The above example is True because London is in England and therefore prompt 2 is clearly implied by prompt 1.)
- *"Russian cosmonaut Valery Polyakov set the record for the longest continuous amount of time spent in space, a staggering 438 days, between 1994 and 1995."*
  "Russians hold record for longest stay in space."
  (The above example is True because the information in the second prompt is contained in the first prompt: Valery is Russian and she set the record for longest stay in space.)
- *"She does not disgree with her brother's opinion, but she believes he's too aggresive in his defense"*
  "She agrees with her brother's opinion, but she believes he's too aggresive in his defense"
  (The above example is True because the second prompt is an exact paraphrase of the first prompt, with exactly the same meaning.)

Choose False if the event or situation described with the first prompt on top does not necessarily imply that this second prompt must also be true. For example,

- *"This method was developed at Columbia and applied to data processing at CERN."*
  "This method was developed at Columbia and applied to data processing at CERN with limited success."
  (The above example is False because the second prompt is introducing new information not implied in the first prompt: The first prompt does not give us any knowledge of how succesful the application of the method at CERN was.)
- *"This building is very tall."*
  "This is the tallest building in New York."
  (The above example is False because a building being tall does not mean it must be the tallest building, nor that it is in New York.)
- *"Hours earlier, Yasser Arafat called for an end to attacks against Israeli civilians in the two weeks before Israeli elections."*
  "Arafat condemned suicide bomb attacks inside Israel."
  (The above example is False because from the first prompt we only know that Arafat called for an end to attacks against Israeli citizens, we do not know what kind of attacks he may have been condemning.)

You do not have to worry about whether the writing style is maintained between the two prompts.

If you have any more questions, please refer to our FAQ page.

Table 9: Task-specific instructions for the Gendered Ambiguous Pronoun Coreference (GAP) task. These instructions were provided during both training and annotation phases.

GAP Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with an extract from a Wikipedia article, with one bolded pronoun. We will also give you two names from the text that this pronoun could refer to. Your job is to figure out, based on the extract, if the pronoun refers to option A, options B, or neither:

Choose <u>A</u> if the pronoun refers to option A. For example,

> *"In 2010 Ella Kabambe was not the official Miss Malawi; this was Faith Chibale, but Kabambe represented the country in the Miss World pageant. At the 2012 Miss World, Susan Mtegha pushed Miss New Zealand, Collette Lochore, during the opening headshot of the pageant, claiming that Miss New Zealand was in her space."*
>
> *Does **her** refer to option A or B below?*

A *Susan Mtegha*

B *Collette Lochore*

C *Neither*

Choose <u>B</u> if the pronoun refers to option B. For example,

> *"In 1650 he started his career as advisor in the ministerium of finances in Den Haag. After he became a minister he went back to Amsterdam, and took place as a sort of chairing mayor of this city. After the death of his brother Cornelis, De Graeff became the strong leader of the republicans. He held this position until the rampjaar."*
>
> *Does **He** refer to option A or B below?*

A *Cornelis*

B *De Graeff*

C *Neither*

Choose <u>C</u> if the pronoun refers to neither option. For example,

> *"Reb Chaim Yaakov's wife is the sister of Rabbi Moishe Sternbuch, as is the wife of Rabbi Meshulam Dovid Soloveitchik, making the two Rabbis his uncles. Reb Asher's brother Rabbi Shlomo Arieli is the author of a critical edition of the novallae of Rabbi Akiva Eiger. Before his marriage, Rabbi Arieli studied in the Ponevezh Yeshiva headed by Rabbi Shmuel Rozovsky, and he later studied under his father-in-law in the Mirrer Yeshiva."*
>
> *Does **his** refer to option A or B below?*

A *Reb Asher*

B *Akiva Eiger*

C *Neither*

If you have any more questions, please refer to our <u>FAQ</u> page.

Table 10: Task-specific instructions for the Paraphrase Adversaries from Word Scrambling (PAWS) task. These instructions were provided during both training and annotation phases.

Paraphrase Detection Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with two similar sentences taken from Wikipedia articles. Your job is to figure out if these two sentences are paraphrases of each other, and convey exactly the same meaning:

Choose Yes if the sentences are paraphrases and have the exact same meaning. For example,

> *"Hastings Ndlovu was buried with Hector Pieterson at Avalon Cemetery in Johannesburg."*
>
> *"Hastings Ndlovu , together with Hector Pieterson , was buried at the Avalon cemetery in Johannesburg ."*
>
> *"The complex of the Trabzon World Trade Center is close to Trabzon Airport ."*
>
> *"The complex of World Trade Center Trabzon is situated close to Trabzon Airport ."*

Choose No if the two sentences are not exact paraphrases and mean different things. For example,

> *"She was only a few months in French service when she met some British frigates in 1809 ."*
>
> *"She was only in British service for a few months , when in 1809 , she encountered some French frigates ."*
>
> *"This work caused him to trigger important reflections on the practices of molecular genetics and genomics at a time when this was not considered ethical ."*
>
> *"This work led him to trigger ethical reflections on the practices of molecular genetics and genomics at a time when this was not considered important ."*

If you have any more questions, please refer to our FAQ page.

Table 11: Task-specific instructions for the Quora Insincere Questions task. These instructions were provided during both training and annotation phases.

Insincere Questions Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with a question that someone posted on Quora. Your job is to figure out whether or not this is a sincere question. An insincere question is defined as a question intended to make a statement rather than look for helpful answers. Some characteristics that can signify that a question is insincere:

- Has a non-neutral tone
  - Has an exaggerated tone to underscore a point about a group of people
  - Is rhetorical and meant to imply a statement about a group of people
- Is disparaging or inflammatory
  - Suggests a discriminatory idea against a protected class of people, or seeks confirmation of a stereotype
  - Makes disparaging attacks/insults against a specific person or group of people
  - Based on an outlandish premise about a group of people
  - Disparages against a characteristic that is not fixable and not measurable
- Isn't grounded in reality
  - Based on false information, or contains absurd assumptions
  - Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

Please note that there are far fewer insincere questions than there are sincere questions! So you should expect to label most questions as sincere.

Examples,

Choose Sincere if you believe the person asking the question was genuinely seeking an answer from the forum. For example,

> *"How do DNA and RNA compare and contrast?"*
> *"Are there any sports that you don't like?"*
> *"What is the main purpose of penance?"*

Choose Insincere if you believe the person asking the question was not really seeking an answer but was being inflammatory, extremely rhetorical, or absurd. For example,

> *"How do I sell Pakistan? I need lots of money so I decided to sell Pakistan any one wanna buy?"*
> *"If Hispanics are so proud of their countries, why do they move out?"*
> *"Why Chinese people are always not welcome in all countries?"*

If you have any more questions, please refer to our FAQ page.

Table 12: Task-specific instructions for the Ultrafine Entity Typing task. These instructions were provided during both training and annotation phases.

Entity Typing Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will provide you with a sentence with on bolded word or phrase. We will also give you a possible tag for this bolded word or phrase. Your job is to decide, in the context of the sentence, if this tag is correct and applicable to the bolded word or phrase:

Choose Yes if the tag is applicable and accurately describes the selected word or phrase. For example,

> *"Spain was the gold line." **It** started out with zero gold in 1937, and by 1945 it had 65.5 tons.*
>
> *Tag: nation*

Choose No if the tag is not applicable and does not describes the selected word or phrase. For example,

> ***Iraqi museum workers** are starting to assess the damage to Iraq's history.*
>
> *Tag: organism*

If you have any more questions, please refer to our FAQ page.

Table 13: Task-specific instructions for the Empathetic Reaction task. These instructions were provided during both training and annotation phases.

Empathy and Distress Analysis Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with a message someone wrote after reading an article. Your job is to figure out, based on this message, how disressed and empathetic the author was feeling. Empathy is defined as feeling warm, tender, sympathetic, moved, or compassionate. Distressed is defined as feeling worried, upset, troubled, perturbed, grieved, distrubed, or alarmed.

Examples,
The author of the following message was not feeling empathetic at all with an empathy score of 1, and was very distressed with a distress score of 7,

> *"I really hate ISIS. They continue to be the stain on society by committing atrocities condemned by every nation in the world. They must be stopped at all costs and they must be destroyed so that they wont hurt another soul. These poor people who are trying to survive get killed, imprisoned, or brainwashed into joining and there seems to be no way to stop them."*

The author of the following message is feeling very empathetic with an empathy score of 7 and also very distressed with a distress score of 7,

> *"All of you know that I love birds. This article was hard for me to read because of that. Wind turbines are killing a lot of birds, including eagles. It's really very sad. It makes me feel awful. I am all for wind turbines and renewable sources of energy because of global warming and coal, but this is awful. I don't want these poor birds to die like this. Read this article and you'll see why."*

The author of the following message is feeling moderately empathetic with an empathy score of 4 and moderately distressed with a distress score of 4,

> *"I just read an article about wild fires sending a smokey haze across the state near the Appalachian mountains. Can you imagine how big the fire must be to spread so far and wide? And the people in the area obviously suffer the most. What if you have asthma or some other condition that restricts your breathing?"*

The author of the following message is feeling very empathetic with an empathy score of 7 and mildly distressed with a distress score of 2,

> *"This is a very sad article. Being of of the first female fighter pilots must have given her and her family great honor. I think that there should be more training for all pilots who deal in these acrobatic flying routines. I also think that women have just as much of a right to become a fighter pilot as men."*

If you have any more questions, please refer to our FAQ page.
