A Conservative Human Baseline Estimate for GLUE: People Still (Mostly) Beat Machines

Nikita Nangia[1]
nikitanangia@nyu.edu

Samuel R. Bowman[1,2,3]
bowman@nyu.edu

[1]Center for Data Science
New York University

[2]Dept. of Linguistics
New York University

[3]Dept. of Computer Science
New York University

Abstract

The GLUE benchmark has seen dramatic progress, with state of the art moving from 69 to over 80 in the past year. Naturally, we wondered if there is any headroom left. In this paper we attempt to provide a fair (if conservative) estimate of human performance on the benchmark using crowdworker annotators. Our human performance baseline reaches a score of 86.9, and we see that humans robustly outperform the current state of the art on six of the nine GLUE tasks. We conduct some analysis and suggest that working with low-resource datasets could be a valuable next step in Natural Language Understanding research.

1 Introduction

Recent prominent work in natural language processing (NLP) has focused on building general purpose models that can learn good language representations across a range of tasks and domains (Peters et al., 2018; McCann et al., 2017; Devlin et al., 2018). The General Language Understanding Evaluation (GLUE; Wang et al., 2019) benchmark was designed to promote the development of models that can handle many different language understanding tasks and genres without needing to be fully retrained on massive amounts of data. GLUE includes nine different Natural Language Understanding (NLU) tasks that test a range of linguistic tasks like natural language inference, sentiment analysis, acceptability judgment, sentence similarity, and common sense reasoning.

The recent Bidirectional Encoder Representations from Transformers (BERT) model by Devlin et al. (2018) is pre-trained with a masked language modeling objective on a large amount of unlabeled data, and then fine-tuned to a specific task. BERT represents state-of-the-art on GLUE, with a wide margin between it and the next best system (GPT on STILTs; Phang et al., 2019; Radford et al., 2018). BERT's performance on GLUE is impressive enough to now prompt the question: How much better are humans at these NLP tasks? Have modern methods exhausted the headroom in typical sentence-level NLU tasks? In the case of some language understanding tasks, like SQuAD 2.0 (Rajpurkar et al., 2018), the current state-of-the-art model, which is built on top of BERT,[1] is extremely close behind human performance baseline. In this work, we estimate human performance on the GLUE test set to see which tasks have substantial remaining headroom between human and machine performance. We also present some analysis and discussion on what direction we think NLP tasks could go in next.

While human performance or interannotator agreement numbers have been reported on some GLUE tasks, the data collection methods used to establish those baselines vary substantially. To maintain consistency in our reported baseline numbers and to ensure that our results are at least roughly comparable to numbers for submitted machine learning models, we collect annotations using a uniform method for all nine GLUE tasks.

To estimate human performance on GLUE, we conduct a data collection effort with crowdworkers, we have detailed this process in Section 3. For each of the nine GLUE tasks, we give crowdworkers a brief training exercise on the task, ask them to annotate a random subset of the evaluation data, and then collect *majority vote* labels from five annotators for each example in the subset. Comparing these labels with the ground-truth GLUE test labels yields an overall estimated GLUE score of 86.9—well above BERT's 80.3—and yields single-task scores that are substantially better than BERT on six of nine tasks. However, given the

[1] https://rajpurkar.github.io/SQuAD-explorer/

	Avg	Single Sentence		Sentence Similarity			Natural Language Inference			
		CoLA	**SST-2**	**MRPC**	**STS-B**	**QQP**	**MNLI**	**QNLI**	**RTE**	**WNLI**
Training Size		*8.5k*	*67k*	*3.7k*	*7k*	*364k*	*393k*	*108k*	*2.5k*	*634*
Human	**86.9**	**66.4**	**97.8**	80.8/86.3	**92.7/92.6**	80.4/59.5	**90.8/91.4**	**91.2**	**93.6**	**95.9**
BERT	80.3	60.5	94.9	**85.4/89.3**	87.6/86.5	**89.3/72.1**	86.7/85.9	**91.1**	70.1	65.1
Δ	6.6	5.9	2.9	-4.6/-3.0	5.1/6.1	-8.9/-12.6	4.1/5.5	0.1	23.5	30.8
BERT-5000	75.8	57.6	92.0	85.4/89.3	87.1/85.8	82.2/61.0	76.4/76.9	89.2	69.2	65.1
BERT-1000	70.7	49.0	90.4	78.5/84.3	83.6/82.3	77.8/55.8	66.5/68.3	86.6	65.6	65.1
BERT-500	68.5	37.2	88.1	74.0/80.7	77.3/75.2	75.4/51.2	61.8/63.0	85.7	61.5	65.1

Table 1: As in the original GLUE paper, we report the Matthews correlation coefficient for CoLA. For MRPC and Quora, we report accuracy and then F1. For STS-B, we report Pearson and then Spearman correlation coefficients. For MNLI, we report accuracy on the matched and then mismatched test sets. For all other tasks we report accuracy. The *Avg* column shows the overall GLUE score: an average across each row, weighting each task equally. The Δ columns shows the difference between the Human performance baseline and BERT. The *Training Size* row gives the size of the full training dataset for each task. The BERT-5000/1000/500 rows show test set results for BERT-Large when it is trained on only 5k, 1k, and 500 examples respectively.

fact that machines, including BERT[2], fail at linguistic tests like the Wug test (Lake and Baroni, 2018), which is designed to gauge acquisition of plural formation rules in children (Berko Gleason, 1958), the performance gap on GLUE is smaller than we expected.

The one striking, and unsurprising, exception is the Winograd Schema NLI Corpus (WNLI; based on Levesque et al., 2012). On this data-poor common-sense reasoning task, humans reach 95.9% accuracy, while no existing machine learning system exceeds the majority-class baseline of 65.1%.

Given the generally impressive performance of BERT on GLUE, we propose that different benchmark NLP tasks for natural language understanding need to be set. We need tasks that challenge machine learning systems in different ways than our current benchmark tasks.

2 Related Work

GLUE (Wang et al., 2019) is comprised of nine tasks, all of which are sentence level classification task. Collectively, the tasks test for different aspects of NLU. MultiNLI (Williams et al., 2018), RTE (competition releases 1–3 and 5, merged and treated as a single binary classification task; Dagan et al. 2006, Bar Haim et al. 2006, Giampiccolo et al. 2007, Bentivogli et al. 2009), and QNLI (an answer sentence selection task based on SQuAD; (Rajpurkar et al., 2016)) test for natural language inference ability. The Microsoft Paraphrase Corpus (MRPC; Dolan and Brockett, 2005), the Semantic Textual Similarity Benchmark (STS-B; Cer et al., 2017), and Quora Question Pairs (QQP)[3] are sentence similarity tasks. The Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2018) tests acceptability judgments. The Stanford Sentiment Treebank (SST; Socher et al., 2013) is a sentiment analysis task. And lastly, WNLI is derived from private data created for the Winograd Schema Challenge (Levesque et al., 2012) and it tests for common sense reasoning.

Warstadt et al. (2018) report human performance numbers on CoLA as well. Using the majority decision from five annotators on 200 examples, they get a Mathhews correlation coefficient (MCC) of 71.3. Their human performance estimate is higher than our finding of a MCC of 66.4. We believe this discrepancy is because Warstadt et al. (2018) use Linguistics PhD students as expert annotators, whereas our crowdworkers are non-experts. This evidence supports our belief that our human performance baseline is a conservative estimate, and that higher performance is possible. Bender (2015) also estimate human performance on the Winograd Schema Challenge. They also use crowdworkers through Amazon's Mechicanical Turk and report an average accuracy of 92.1%. Their finding is 4 points lower than ours, this difference may be a factor of data collection methods and worker pool.

In their paper introducing the Natural Questions corpus, Kwiatkowski et al. (2019) measure an up-

[2]This claim is based on an experiment by Yoav Goldberg where BERT is tested on the Wug test: https://tinyurl.com/yd8gxrv2

[3]https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

per bound on the task with a 25-way annotated subset of the corpus. While this work on human performance baselines is similar, we only use five annotators and do not believe our baseline is true upper bounds for the tasks. Rather than focusing on a single dataset, this work establishes conservative baselines for human performance across nine GLUE tasks using consistent experimental methods.

3 Data Collection Method

To establish human performance on GLUE tasks, we hire annotators through the Hybrid[4] data collection platform, which is similar to Amazon's Mechanical Turk. We conduct the data collection in two phases: Each worker first completes a short training procedure then moves on to the main annotation task. For the annotation, we tune the pay rate by task since the average time taken per example varies by task. The average pay rate is $17/hour.

Training In the training phase for each GLUE task, the workers answers 20 randomly sampled examples from the task development set. On the training page they are linked to instructions that are tailored to each task. On each page of the worker training, five examples are shown and the answers can be revealed by clicking on a "Show" button at the bottom of the page. The workers are instructed to answer each set of questions and check their work so they can familiarize themselves with the task. Workers who get less than 65% of the examples correct during training are not allowed to qualify for the main task. This is an arbitrary, and intentionally low, threshold. The purpose of this step is to train the workers on the task, not to weed out bad workers. With our data-collection framework, it is possible for workers to change their answers after viewing the correct labels, so we can not use the training phase a as a serious filter. (See Appendix A.1 for details on the training phase.)

Annotation Upon finishing training, the workers move onto the annotation phase, which is our source for human performance baseline on GLUE. For this data collection we randomly sample 500 examples from each task's test set, with the exception of WNLI where we sample 145 of the 147 available test set examples (the two missing examples are the result of a data preparation error). For each of these sampled data points, we collect five annotations from five different workers (see Appendix A.2) We use the test set since the test and development sets are qualitatively different for some tasks, and since we wish compare our results directly with those on the GLUE leaderboard.[5]

4 Results

To calculate the human performance baseline, we take the majority vote across the 5 crowd-sourced annotations. In the case of MultiNLI, since there are three possible labels—*entailment*, *neutral*, and *contradiction*—there are a few ties ($\sim$2%). In these two-way ties, we take the label that is more frequent in the development set. And in the case of STS-B, we take a average of the scalar annotator labels. Since we only collected annotations for a subset of the data, we could not submit our results to the GLUE leaderboard. Instead, we worked in cooperation with the GLUE team to get scores.

The human performance scores are shown in the first line of Table 1. These results shows that our annotators outperform BERT overall on GLUE. However, on five of the tasks, the gap is less than 6 points wide. And on MRPC and QQP, the BERT machine outperforms our annotators by a considerable margin.

The results on QQP are particularly surprising, BERT scores 12.6 F1 points more than our annotators. Our annotators however, are only given 20 examples and a short set of instructions to train them on GLUE tasks. By comparison, BERT is trained on 364k QQP examples. We know that people are very efficient learners but machine learning systems still struggle with adapting to new tasks and domains in data-poor settings. This may be further evidenced by BERT's performance discrepancy between MultiNLI and RTE. Both MultiNLI and RTE are textual entailment datasets, they primarily differ in their data sources. MultiNLI comes with 393k training examples, while the GLUE version of the RTE task has only 2.5k examples, and this is likely the source of the 15 point disparity in scores. In comparison, human performance on MultiNLI and RTE are within two points of one another.

To take a closer look at BERT's sample complexity, we train it on 5k, 1k, and 500 examples

[4] http://www.gethybrid.io

[5] https://gluebenchmark.com/leaderboard

for each GLUE task (or fewer for tasks with fewer training examples). We use the publicly available implementation of BERT released by Devlin et al. (2018). We use the publicly distributed pre-trained weights as the initialization point for training on the GLUE tasks. The results are shown in the last two lines of Table 1. We see a precipitous drop in performance on most tasks with large datasets. One exception here is QNLI, which may be because, for QNLI, there is minimal domain shift for BERT since BERT is trained on English Wikipedia, and the proposed answers in QNLI are spans from Wikipedia (Rajpurkar et al., 2016; Wang et al., 2019). On MRPC and QQP however, BERT's performance drops below human performance in the 1k and 500-example low-data setting.

5 Discussion

Our conservative estimate of human performance shows that our annotators beat state-of-the-art BERT on GLUE overall by 6.6 points. While the human baseline is better than BERT on six of nine individual tasks, our results suggest that there isn't a lot of headroom left in the current GLUE framework, with the exception of the WNLI task.

No system on the GLUE leaderboard has managed to exceed the performance of the most-frequent-class baseline on WNLI, and several papers that propose methods for GLUE justify their poor performance by asserting that the task must be somewhat broken.[6] While WNLI was constructed so as not to include any statistical cues that a simple machine learning system can exploit, which can make it quite difficult, the WNLI test set nonetheless shows one of the *highest* human performance scores of the nine GLUE tasks, reflecting its status as a corpus constructed and vetted by artificial intelligence experts. This makes it clear that tasks like WNLI with small training sets (634 sentence pairs) and no simple cues remain a serious (and sometimes unacknowledged) blind spot for modern neural network sentence understanding methods.

We do find that BERT outperforms the human baseline on MRPC and QQP. A qualitative analysis of the examples that our annotators get wrong on MRPC and QQP shows that the labels in these instances often do not match the colloquial meaning of paraphrase. It is possible that machine learning systems are able to pick up on some aggregate statistics that our annotators don't have access to. It is also possible that with more training human annotators will be able to match machine performance on these two tasks.

In experimenting with BERT trained on 1k, 5k, and 500 examples, we see that BERT suffers in low-resource settings. This result gives us more reason to believe that low-resource settings continue to be challenging for machine learning systems. If we want more robust, flexible, and easily adaptable machine systems, designing them to have low sample complexity will be a step in the right direction.

6 Conclusion

This paper presents a conservative estimate of human performance to serve as a performance target for the GLUE sentence understanding benchmark. We obtain this baseline with the help of crowdworker annotators. We see that state-of-the-art models like BERT are not far behind human performance on most GLUE tasks, but we also note that, when trained in low-resource settings, BERT's performance falls considerably. Given these results, and the continued difficulty neural methods have with the Winograd Schema Challenge, we propose that future NLU benchmark datasets could provide valuable challenges by being data-poor.

Acknowledgments

This project has benefited from financial support to Sam Bowman from Samsung Research. We thank Alex Wang and Amanpreet Singh for their help with conducting GLUE evaluations, and we thank Jason Phang for his help with training the BERT model.

References

Roy Bar Haim, Ido Dagan, Bill Dolan, Ferro Lisa, 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*.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *MAICS*.

Luisa Bentivogli, Ido Dagan, Hao Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The

[6] Devlin et al. (2018), for example, mention that they avoid "the problematic WNLI set".

fifth PASCAL recognizing textual entailment challenge. In *Proceedings of Text Analysis Conference*.

Jean Berko Gleason. 1958. The child's learning of english morphology. *Word*, 14.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. 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. Association for Computational Linguistics.

Ido Dagan, Oren Glickmen, and Magnini Bernardo. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint 1810.04805*.

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

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research.

Brenden M. Lake and Marco Baroni. 2018. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of 35th International Conference on Machine Learning*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint 1811.01088*.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. 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.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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.

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 someone spoke. Your job is to figure out, based on this sentence, if the speaker is a native speaker of English. You should ignore the general topic of the sentence and focus on the fluency of the sentence.

- Choose correct if you think the sentence sounds fluent and you think it was spoken by a native-English speaker. Examples:

  "A hundred men surrounded the fort.

  "Everybody who attended last weeks huge rally, whoever they were, signed the petition."

  "Where did you go and who ate what?"

- Choose incorrect if you think the sentence does not sound completely fluent and may have been spoken by a non-native English speaker. Examples:

  "Sue gave to Bill a book.

  "Mary came to be introduced by the bartender and I also came to be."

  "The problem perceives easily."

Table 2: The instructions given to crowd-sourced worker for the CoLA task. While the instructions were tailored for each task in GLUE, they all followed a similar format.

A Crowd-Sourced Data Collection

A.1 Training Phase

During training, we provide a link to task-specific instructions. As an example, the instructions for CoLA are shown in Table 2. The instructions for all tasks follow the same format: briefly describing the annotator's job, explaining the labels, and providing at least one example. We've included a zip file with the instructions for all the GLUE tasks.

In addition to the task-specific instruction, we also provide general instructions on the training phase. An example of these instructions is shown in Table 3. The only variation from task to task, is the name of task in the instructions. Lastly, we provide a link to an FAQ page. The FAQ page addresses the balance of the data. If the labels are balanced, we tell the annotators so. If the labels are not balanced, we assure the annotators that they need not worry about assigning one label more frequently. For most tasks we also inform the annotators where the data comes from, for example from another crowdsourcing effort or from news articles.

During training, each page is headed by the instructions and then the annotator is given five examples to label. At the bottom of the page, there is a "Show" button which reveals the answers. If their submitted answer was incorrect, the correct label is shown in red, otherwise it's in black. In the instructions, the worker is asked to check their work with this button.

A.2 Annotation Phase

In the main data collection phase we simply provide the annotator a link to the same task-specific instructions used in the training phase (Figure 2). We also provide a link to the same FAQ page as in the training phase. We enforce the training phase as a qualification for annotation, so crowdworkers can not participate in annotation without first completing the associated training.

This project is a training task that needs to be completed before working on the main project on Hybrid named Human Performance: CoLA. For this CoLA task, we have the true label and we want to get information on how well people do on the task. This training is short but is designed to help you get a sense of the questions and the expected labels.

Please note that the pay per HIT for this training task is also lower than it is for the main project Human Performance: CoLA. Once you are done with the training, please proceed to the main task!

In this training, you must answer all the questions on the page and then, to see how you did, click the Show button at the bottom of the page before moving onto the next HIT. The Show button will reveal the true labels. If you answered correctly, the revealed label will be in black, otherwise it will be in red. Please use this training and the provided answers to build an understanding of what the answers to these questions looks like (the main project, Human Performance: CoLA , does not have the answers on the page).

Table 3: Instructions about the training phase provided to workers. This example is for CoLA training. The only change in instructions for other tasks is the name of the task.