GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

Alex Wang,[1] Amanpreet Singh,[1] Julian Michael,[2] Felix Hill,[3] Omer Levy,[2] and Samuel R. Bowman[1]

[1]New York University, New York, NY
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA
[3]DeepMind, London, UK
{alexwang,amanpreet,bowman}@nyu.edu
{julianjm,omerlevy}@cs.washington.edu
felixhill@google.com

Human ability to understand language is *general*, *flexible*, and *robust*. In contrast, most NLU models above the word level are designed for a specific task and struggle with out-of-domain data. If we aspire to develop models with understanding beyond the detection of superficial correspondences between inputs and outputs, then it is critical to develop a unified model that can execute a range of linguistic tasks across different domains.

To facilitate research in this direction, we present the General Language Understanding Evaluation (GLUE, gluebenchmark.com): a benchmark of nine diverse NLU tasks, an auxiliary dataset for probing models for understanding of specific linguistic phenomena, and an online platform for evaluating and comparing models. For some benchmark tasks, training data is plentiful, but for others it is limited or does not match the genre of the test set. GLUE thus favors models that can represent linguistic knowledge in a way that facilitates sample-efficient learning and effective knowledge-transfer across tasks. While none of the datasets in GLUE were created from scratch for the benchmark, four of them feature privately-held test data, which is used to ensure that the benchmark is used fairly.

We evaluate baselines that use ELMo (Peters et al., 2018), a powerful transfer learning technique, as well as state-of-the-art sentence representation models. The best models still achieve fairly low absolute scores. Analysis with our diagnostic dataset yields similarly weak performance over all phenomena tested, with some exceptions.

The GLUE benchmark GLUE consists of nine English sentence understanding tasks covering a broad range of domains, data quantities, and difficulties. As the goal of GLUE is to spur development of generalizable NLU systems, we design the benchmark such that good performance should require models to share substantial knowledge (e.g., trained parameters) across tasks, while maintaining some task-specific components. Though it is possible to train a model per task and evaluate the resulting set of models on this benchmark, we expect that inclusion of several data-scarce tasks will ultimately render this approach uncompetitive.

Corpus	|Train|	Task	Domain
Single-Sentence Tasks			
CoLA	8.5k	acceptability	misc.
SST-2	67k	sentiment	movie reviews
Similarity and Paraphrase Tasks			
MRPC	3.7k	paraphrase	news
STS-B	7k	textual sim.	misc.
QQP	364k	paraphrase	online QA
Inference Tasks			
MNLI	393k	NLI	misc.
QNLI	108k	QA/NLI	Wikipedia
RTE	2.5k	NLI	misc.
WNLI	634	coref./NLI	fiction books

Table 1: Task descriptions and statistics. **Bold** denotes tasks for which there is privately-held test data. All tasks are binary classification, except STS-B (regression) and MNLI (three classes).

The nine tasks include two tasks with single-sentence inputs: Corpus of Linguistic Acceptability (CoLA; Warstadt et al. 2018) and Stanford Sentiment Treebank (SST-2; Socher et al. 2013) Three tasks involve detecting semantic similarity: Microsoft Research Paraphrase Corpus (MRPC, (Dolan and Brockett, 2005)), Quora Question Pairs[1] (QQP), and Semantic Textual Similarity Benchmark (STS-B; Cer et al. 2017). The remaining four tasks are formatted as natural language inference (NLI) tasks, such as the Multi-Genre NLI corpus (MNLI; Williams et al. 2018) and Recog-

[1] data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355

Model	Avg	Single Sentence CoLA	SST-2	Similarity and Paraphrase MRPC	QQP	STS-B	Natural Language Inference MNLI	QNLI	RTE	WNLI
Single-task	64.8	**35.0**	90.2	68.8/80.2	**86.5/66.1**	55.5/52.5	**76.9/76.7**	61.1	50.4	**65.1**
Multi-task	**69.0**	18.9	**91.6**	**77.3/83.5**	85.3/63.3	72.8/71.1	75.6/75.9	81.7	**61.2**	**65.1**
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	75.1	54.1	62.3
Skip-Thought	61.5	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	74.7	53.1	**65.1**
InferSent	64.7	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	79.8	58.0	**65.1**
DisSent	62.1	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	75.2	56.4	**65.1**
GenSen	66.6	7.7	83.1	76.6/83.0	82.9/59.8	**79.3/79.2**	71.4/71.3	**82.3**	59.2	**65.1**

Table 2: Baseline performance on the GLUE tasks. For MNLI, we report accuracy on the matched and mismatched test sets. For MRPC and QQP, we report accuracy and F1. For STS-B, we report Pearson and Spearman correlation. For CoLA, we report Matthews correlation (Matthews, 1975). For all other tasks we report accuracy. All values are scaled by 100. A similar table is presented on the online platform.

nizing Textual Entailment (RTE; aggregated from Dagan et al. 2006, Bar Haim et al. 2006, Giampiccolo et al. 2007, Bentivogli et al. 2009), as well as versions of SQuAD (Rajpurkar et al., 2016) and Winograd Schema Challenge (Levesque et al., 2011) recast as NLI (resp. QNLI, WNLI). Table 1 summarizes the tasks. Performance on the benchmark is measured per task as well as in aggregate, averaging performance across tasks.

Diagnostic Dataset To understand the types of knowledge learned by models, GLUE also includes a dataset of hand-crafted examples for probing trained models. This dataset is designed to highlight common phenomena, such as the use of world knowledge, logical operators, and lexical entailments, that models must grasp if they are to robustly solve the tasks. Each of the 550 examples is an NLI sentence pair tagged with the phenomena demonstrated. We ensure that the data is reasonably diverse by producing examples for a wide variety of linguistic phenomena, and basing our examples on naturally-occurring sentences from several domains. We validate our data by using the hypothesis-only baseline from Gururangan et al. (2018) and having six NLP researchers manually validate a random sample of the data.

Baselines To demonstrate the benchmark in use, we apply multi-task learning on the training data of the GLUE tasks, via a model that shares a BiLSTM between task-specific classifiers. We also train models that use the same architecture but are trained on a single benchmark task. Finally, we evaluate the following pretrained models: average bag-of-words using GloVe embeddings (CBoW), Skip-Thought (Kiros et al., 2015), InferSent (Conneau et al., 2017), DisSent (Nie et al., 2017), and GenSen (Subramanian et al., 2018).

Tags	Sentence Pair
Quantifiers *Double Negation*	I have never seen a hummingbird not flying. I have never seen a hummingbird.
Active/Passive	Cape sparrows eat seeds, along with soft plant parts and insects. Cape sparrows are eaten.
Named Entities *World Knowledge*	Musk decided to offer up his personal Tesla roadster. Musk decided to offer up his personal car.

Table 3: Diagnostic set examples. Systems must predict the relationship between the sentences, either *entailment*, *neutral*, or *contradiction* when one sentence is the premise and the other is the hypothesis, and vice versa. Examples are tagged with the phenomena demonstrated. We group each phenomena into one of four broad categories.

We find that our models trained directly on the GLUE tasks generally outperform those that do not, though all models obtain fairy low absolute scores. Probing the baselines with the diagnostic data, we find that performance on the benchmark correlates with performance on the diagnostic data, and that the best baselines similarly achieve low absolute performance on the linguistic phenomena included in the diagnostic data.

Conclusion We present the GLUE benchmark, consisting of: (i) a suite of nine NLU tasks, built on established annotated datasets and covering a diverse range of text genres, dataset sizes, and difficulties; (ii) an online evaluation platform and leaderboard, based primarily on private test data; (iii) an expert-constructed analysis dataset. Experiments indicate that solving GLUE is beyond the capability of current transfer learning methods.

References

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

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

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *11th International Workshop on Semantic Evaluations*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 681–691.

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

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

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

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning*, volume 46, page 47.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Allen Nie, Erin D Bennett, and Noah D Goodman. 2017. Dissent: Sentence representation learning from explicit discourse relations. *arXiv preprint 1710.04334*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL 2018*.

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*, pages 1631–1642.

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

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

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL 2018*.