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

Abstract

For natural language understanding (NLU) technology to be maximally useful, both practically and as a scientific object of study, it must be general: it must be able to process language in a way that is not exclusively tailored to any one specific task or dataset. In pursuit of this objective, we introduce the General Language Understanding Evaluation benchmark (GLUE), a tool for evaluating and analyzing the performance of models across a diverse range of existing NLU tasks. GLUE is model-agnostic, but it incentivizes sharing knowledge across tasks because certain tasks have very limited training data. We further provide a hand-crafted diagnostic test suite that enables detailed linguistic analysis of NLU models. We evaluate baselines based on current methods for multi-task and transfer learning and find that they do not immediately give substantial improvements over the aggregate performance of training a separate model per task, indicating room for improvement in developing general and robust NLU systems.

1 Introduction

Human ability to understand language is *general*, *flexible*, and *robust*. We can effectively interpret and respond to utterances of diverse form and function in many different contexts. In contrast, most natural language understanding (NLU) models above the word level are designed for one particular task and struggle with out-of-domain data. If we aspire to develop models whose understanding extends beyond the detection of superficial correspondences between inputs and outputs, then it is critical to understand how a single model can learn to execute a range of different linguistic tasks on language from different domains.

To motivate research in this direction, we present the General Language Understanding Evaluation benchmark (GLUE, gluebenchmark.com), an online tool for evaluating the performance of a single NLU model across multiple tasks, including question answering, sentiment analysis, and textual entailment, built largely on established existing datasets. GLUE does not place any constraints on model architecture beyond the ability to process single-sentence and paired-sentence inputs and to make corresponding predictions. For some GLUE tasks, directly pertinent training data is plentiful, but for others, training data is limited or fails to match the genre of the test set. GLUE therefore favors models that can learn to represent linguistic and semantic knowledge in a way that facilitates sample-efficient learning and effective knowledge transfer across tasks.

Though GLUE is designed to prefer models with general and robust language understanding, we cannot entirely rule out the existence of simple superficial strategies for solving any of the included tasks. We therefore also provide a set of newly constructed evaluation data for the analysis of model performance. Unlike many test sets employed in machine learning research that reflect the frequency distribution of naturally occurring data or annotations, this dataset is designed to highlight points of difficulty that are relevant to model development and training, such as the incorporation of world knowledge, or the handling of lexical entailments and negation. Visitors to the online platform have access to a breakdown of how well each model handles these phenomena alongside its scores on the primary GLUE test sets.

To better understand the challenged posed by the GLUE benchmark, we conduct experiments with simple baselines and state-of-the-art models for sentence representation. We find that naïve multi-task learning with standard models over the available task training data yields overall perfor-

Corpus	|Train|	|Dev|	|Test|	Task	Metric	Domain
			Single-Sentence Tasks			
CoLA	10k	1k	1.1k	acceptability	Matthews	linguistics literature
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks			
MRPC	4k	N/A	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman	misc.
QQP	400k	N/A	391k	paraphrase	acc./F1	social QA Questions
			Inference Tasks			
MNLI	393k	20k	20k	NLI	acc. (match/mismatch)	misc.
QNLI	108k	11k	11k	QA/NLI	acc.	Wikipedia
RTE	2.7k	N/A	3k	NLI	acc.	misc.
WNLI	706	N/A	146	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-Benchmark, which is a regression task. MNLI has three classes while all other classification tasks are binary.

mance no better than can be achieved by training on a separate model for each task, indicating the need for improved general NLU systems. However, for certain tasks with less training data, we find that multi-task learning approaches do improve over a single-task model. This indicates that there is potentially interesting space for meaningful knowledge sharing across NLU tasks. Analysis with our diagnostic dataset reveals that current models deal well with strong lexical signals and struggle with logic, and that there are interesting patterns in the generalization behavior of our models that do not correlate perfectly with performance on the main benchmark.

In summary, we offer the following contributions:

- A suite of nine sentence- or sentence-pair NLU tasks, built on established annotated datasets where possible, and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty.

- An online evaluation platform and leaderboard, based primarily on privately-held test data. The platform is model-agnostic; any model or method capable of producing results on all nine tasks can be evaluated.

- A suite of diagnostic evaluation data aimed to give model developers feedback on the types of linguistic phenomena their evaluated systems handle well.

- Results with several major existing sentence representation systems such as Skip-Thought (Kiros et al., 2015), InferSent (Conneau et al., 2017), DisSent (Nie et al., 2017), and GenSen (Subramanian et al., 2018).

2 Related Work

Our work builds on various strands of NLP research that aspired to develop better general understanding in models.

Multi-task Learning in NLP Multi-task learning has a rich history in NLP as an approach for learning more general language understanding systems. Collobert et al. (2011), one of the earliest works exploring deep learning for NLP, used a multi-task model to jointly learn POS tagging, chunking, named entity recognition, and semantic role labeling. More recently, there has been work using labels from core NLP tasks to supervise training of lower levels of deep neural networks (Søgaard and Goldberg, 2016; Hashimoto et al., 2016) and automatically learning cross-task sharing mechanisms for multi-task learning (Ruder et al., 2017).

Evaluating Sentence Representations Beyond multi-task learning, much of the work so far towards developing general NLU systems has focused on the development of sentence-to-vector encoder functions (Le and Mikolov, 2014; Kiros et al., 2015, i.a.), including approaches leveraging unlabeled data (Hill et al., 2016; Peters et al., 2018), labeled data (Conneau and Kiela, 2018; McCann et al., 2017), and combinations of these (Collobert et al., 2011; Subramanian et al., 2018).

In this line of work, a standard evaluation practice has emerged, and has recently been codified as SentEval (Conneau et al., 2017; Conneau and Kiela, 2018). Like GLUE, SentEval also relies on a variety of existing classification tasks that involve either one or two sentences as inputs, but only evaluates sentence-to-vector encoders. Specifically, SentEval takes a pre-trained sentence encoder as input and feeds its output encodings into lightweight task-specific models (typically linear classifiers) that are trained and tested on task-specific data.

SentEval is well-suited for evaluating general-purpose sentence representations *in isolation*. However, cross-sentence contextualization and alignment, such as that yielded by methods like soft attention, is instrumental in achieving state-of-the-art performance on tasks such as machine translation (Bahdanau et al., 2014; Vaswani et al., 2017), question answering (Seo et al., 2016; Xiong et al., 2016), and natural language inference[1]. GLUE is designed to facilitate the development of these methods: it is model-agnostic, allowing for any kind of representation or contextualization, including models that use no systematic vector or symbolic representations for sentences whatsoever.

GLUE also diverges from SentEval in the selection of evaluation tasks that are included in the suite. Many of the SentEval tasks are closely related to sentiment analysis, with the inclusion of MR (Pang and Lee, 2005), SST (Socher et al., 2013), CR (Hu and Liu, 2004), and SUBJ (Pang and Lee, 2004). Other tasks are so close to being solved that evaluation on them is less informative, such as MPQA (Wiebe et al., 2005) and TREC (Voorhees et al., 1999). In GLUE, we have attempted to construct a benchmark that is diverse, spans multiple domains, and is systematically difficult.

Evaluation Platforms and Competitions in NLP Our use of an online evaluation platform with private test labels is inspired by a long tradition of shared tasks at the SemEval (Agirre et al., 2007) and CoNLL (Ellison, 1997) conferences, as well as similar leaderboards on Kaggle and CodaLab. These frameworks tend to focus on a single task, while GLUE emphasizes the need to perform well on multiple different tasks using shared model components.

Weston et al. (2015) similarly proposed a hierarchy of tasks towards building question answering and reasoning models, although involving synthetic language, whereas almost all of our data is human-generated. The recently proposed dialogue systems framework ParlAI (Miller et al., 2017) also combines many language understanding tasks into a single framework, although this aggregation is very flexible, and the framework includes no standardized evaluation suite for system performance.

3 Tasks

We aim for GLUE to spur development of generalizable NLU systems. As such, we expect that doing well on the benchmark should require a model to share substantial knowledge (e.g. in the form of trained parameters) across all tasks, while keeping the task-specific components as minimal as possible. Though it is possible to train a single model for each task and evaluate the resulting set of models on this benchmark, we expect that for some data-scarce tasks in the benchmark, knowledge sharing between tasks will be necessary for competitive performance. In such a case, a more unified approach should prevail.

The GLUE benchmark consists of nine English sentence understanding tasks selected to cover a broad spectrum of task type, domain, amount of data, and difficulty. We describe them here and provide a summary in Table 1. Unless otherwise mentioned, tasks are evaluated on accuracy and have a balanced class split.

The benchmark follows the same basic evaluation model of SemEval and Kaggle. To evaluate a system on the benchmark, one must configure that system to perform all of the tasks, run the system on the provided test data, and upload the results to the website for scoring. The site will then show the user (and the public, if desired) an overall score for the main suite of tasks, and per-task scores on both the main tasks and the diagnostic dataset.

3.1 Single-Sentence Tasks

CoLA The Corpus of Linguistic Acceptability[2] consists of examples of expert English sentence acceptability judgments drawn from 22 books and

[1] In the case of SNLI (Bowman et al., 2015), the best-performing sentence encoding model on the leaderboard as of April 2018 achieves 86.3% accuracy, while the best performing attention-based model achieves 89.3%.

[2] Available at: `nyu-mll.github.io/CoLA`

journal articles on linguistic theory. Each example is a single string of English words annotated with whether it is a grammatically possible sentence of English. Superficially, this data is similar to our analysis data in that it is constructed to demonstrate potentially subtle and difficult contrasts. However, judgments of this particular kind are the primary form of evidence in linguistic theory (Schütze, 1996), and were a machine learning system to be able to predict them reliably, it would offer potentially substantial evidence on questions of language learnability and innate bias. As in MNLI, the corpus contains development and test examples drawn from in-domain data (the same books and articles used in the training set) and out-of-domain data, though we report numbers only on the unified development and test sets without differentiating these. We follow the original work and report the Matthews correlation coefficient (Matthews, 1975), which evaluates classifiers on unbalanced binary classification tasks with a range from -1 to 1, with 0 being the performance at random chance. We use the standard test set, for which we obtained labels privately from the authors.

SST-2 The Stanford Sentiment Treebank (Socher et al., 2013) consists of sentences extracted from movie reviews and human annotations of their sentiment. Given a sentence, the task is to determine the sentiment of the sentence. We use the two-way (positive/negative) class split.

3.2 Similarity and Paraphrase Tasks

MRPC The Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005) is a corpus of sentence pairs automatically extracted from online news sources, with human annotations of whether the sentences in the pair are semantically equivalent. Because the classes are imbalanced (68% positive, 32% negative), we follow common practice and report both accuracy and F1 score.

QQP The Quora Question Pairs[3] dataset is a collection of question pairs from the community question-answering website Quora. Given two questions, the task is to determine whether they are semantically equivalent. As in MRPC, the class distribution in QQP is unbalanced (37% positive, 63% negative), so we report both accuracy and F1 score. We use the standard test set, for which we obtained labels privately from the authors.

STS-B The Semantic Textual Similarity Benchmark (Cer et al., 2017) is based on the datasets for a series of annual challenges for the task of determining the similarity on a continuous scale from 1 to 5 of a pair of sentences drawn from various sources. We use the STS-Benchmark release, which draws from news headlines, video and image captions, and natural language inference data, scored by human annotators. We follow common practice and evaluate using Pearson and Spearman correlation coefficients between predicted and ground-truth scores.

3.3 Inference Tasks

MNLI The Multi-Genre Natural Language Inference Corpus (Williams et al., 2018) is a crowd-sourced collection of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis, contradicts the hypothesis, or neither (*neutral*). The premise sentences are gathered from a diverse set of sources, including transcribed speech, popular fiction, and government reports. The test set is broken into two sections: *matched*, which is drawn from the same sources as the training set, and *mismatched*, which uses different sources and thus requires domain transfer. We use the standard test set, for which we obtained labels privately from the authors, and evaluate on both sections.

Though not part of the benchmark, we use and recommend the Stanford Natural Language Inference corpus (Bowman et al. 2015; SNLI) as auxiliary training data. It is distributed in the same format for the same task, and has been used productively in cotraining for MNLI (Chen et al., 2017; Gong et al., 2018).

QNLI The Stanford Question Answering Dataset (Rajpurkar et al. 2016; SQuAD) is a question-answering dataset consisting of question-paragraph pairs, where the one of the sentences in the paragraph (drawn from Wikipedia) contains the answer to the corresponding question (written by an annotator). We automatically convert the original SQuAD dataset into a sentence pair classification task by forming a pair between a question and each sentence in the corresponding context. The task is to determine whether the context sentence contains the answer

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

to the question. We filter out pairs where there is low lexical overlap[4] between the question and the context sentence. Specifically, we select all pairs in which the most similar sentence to the question was *not* the answer sentence, as well as an equal amount of cases in which the correct sentence was the most similar to the question, but another distracting sentence was a close second. This approach to converting pre-existing datasets into NLI format is closely related to recent work by White et al. (2017) as well as to the original motivation for textual entailment presented by Dagan et al. (2006). Both argue that many NLP tasks can be productively reduced to textual entailment. We call this processed dataset QNLI (Question-answering NLI).

RTE The Recognizing Textual Entailment (RTE) datasets come from a series of annual challenges for the task of textual entailment, also known as NLI. We combine the data from RTE1 (Dagan et al., 2006), RTE2 (Bar Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009)[5]. Each example in these datasets consists of a premise sentence and a hypothesis sentence, gathered from various online news sources. The task is to predict if the premise entails the hypothesis. We convert all the data to a two-class split (*entailment* or *not entailment*, where we collapse *neutral* and *contradiction* into *not_entailment* for challenges with three classes) for consistency.

WNLI The Winograd Schema Challenge (Levesque et al., 2011) is a reading comprehension challenge where each example consists of a sentence containing a pronoun and a list of its possible referents in the sentence. The task is to determine the correct referent. The data is designed to foil simple statistical methods; it is constructed so that each example hinges on contextual information provided by a single word or phrase in the sentence, which can be switched out to change the answer. We use a small evaluation set consisting of new examples derived from fiction books[6] that was shared privately by the authors of the corpus. To convert the problem into a sentence pair classification task, we construct two sentence pairs per example by replacing the ambiguous pronoun with each possible referent. The task (a slight relaxation of the original Winograd Schema Challenge) is to predict if the sentence with the pronoun substituted is entailed by the original sentence. While the included training set is balanced between two classes (entailment and not entailment), the test set is imbalanced between them (35% entailment, 65% not entailment). We call the resulting sentence pair version of the dataset WNLI (Winograd NLI).

3.4 Scoring

In addition to each task's metric or metrics, Our benchmark reports a macro-average of the metrics over all tasks (see Table 5) to determine a system's position on the leaderboard. For tasks with multiple metrics (e.g., accuracy and F1), we use unweighted average of the metrics as the score for the task.

3.5 Data and Bias

The tasks listed above are meant to represent a diverse sample of those studied in contemporary research on applied sentence-level language understanding, but we do not endorse the use of the task training sets for any specific non-research application. They do not cover every dialect of English one may wish to handle, nor languages outside of English, and as all of them contain text or annotations that were collected in uncontrolled settings, they contain evidence of stereotypes and biases that one may not wish their system to learn (Rudinger et al., 2017).

4 Diagnostic Dataset

Drawing inspiration from the FraCaS test suite (Cooper et al., 1996) and the recent Build-It-Break-It competition (Ettinger et al., 2017), we include a small, manually-curated test set to allow for fine-grained analysis of system performance on a broad range of linguistic phenomena. While the main benchmarks mostly reflect an application-driven distribution of examples (e.g. the question answering dataset will contain questions that people are likely to ask), our diagnostic dataset is collected to highlight a pre-defined set of modeling-relevant phenomena.

Specifically, we construct a set of NLI examples with fine-grained annotations of the linguistic phe-

[4]To measure lexical overlap we use a CBoW representation with pre-trained GloVe embeddings.

[5]RTE4 is not publicly available, while RTE6 and RTE7 do not fit the standard NLI task.

[6]See similar examples at `cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html`

LS	PAS	L	K	Sentence 1	Sentence 2	Fwd	Bwd
	✓			Cape sparrows eat seeds, along with soft plant parts and insects.	Seeds, along with soft plant parts and insects, are eaten by cape sparrows.	E	E
	✓			Cape sparrows eat seeds, along with soft plant parts and insects.	Cape sparrows are eaten by seeds, along with soft plant parts and insects.	N	N
✓	✓			Tulsi Gabbard disagrees with Bernie Sanders on what is the best way to deal with Bashar al-Assad.	Tulsi Gabbard and Bernie Sanders disagree on what is the best way to deal with Bashar al-Assad.	E	E
✓			✓	Musk decided to offer up his personal Tesla roadster.	Musk decided to offer up his personal car.	E	N
			✓	The announcement of Tillerson's departure sent shock waves across the globe.	People across the globe were not expecting Tillerson's departure.	E	N
			✓	The announcement of Tillerson's departure sent shock waves across the globe.	People across the globe were prepared for Tillerson's departure.	C	C
		✓		I have never seen a hummingbird not flying.	I have never seen a hummingbird.	N	E
	✓			Understanding a long document requires tracking how entities are introduced and evolve over time.	Understanding a long document requires evolving over time.	N	N
		✓		Understanding a long document requires tracking how entities are introduced and evolve over time.	Understanding a long document requires understanding how entities are introduced.	E	N
✓				That perspective makes it look gigantic.	That perspective makes it look minuscule.	C	C

Table 2: Examples from the analysis set. Sentence pairs are labeled according to four coarse categories: *Lexical Semantics* (**L**), *Predicate-Argument Structure* (**PAS**), *Logic* (**L**), and *Knowledge and Common Sense* (**K**). Within each category, each example is also tagged with fine-grained labels (see tables 4). See `gluebenchmark.com` for details on the set of labels, their meaning, and how we do the categorization.

nomena they capture. The NLI task is well suited to this kind of analysis, as it is constructed to make it straightforward to evaluate the full set of skills involved in (ungrounded) sentence understanding, from the resolution of syntactic ambiguity to pragmatic reasoning with world knowledge. We ensure that the examples in the diagnostic dataset have a reasonable distribution over word types and topics by building on naturally-occurring sentences from several domains. Table 2 shows examples from the dataset.

Linguistic Phenomena We tag every example with fine- and coarse-grained categories of the linguistic phenomena they involve (categories shown in Table 3). While each example was collected with a single phenomenon in mind, it is often the case that it falls under other categories as well. We therefore code the examples under a non-exclusive tagging scheme, in which a single example can participate in many categories at once. For example, to know that *I like some dogs* entails *I like some animals*, it is not sufficient to know that *dog* lexically entails *animal*; one must also know that *dog/animal* appears in an upward monotone context in the sentence. This example would be classified under both *Lexical Semantics* > *Lexical Entailment* and *Logic* > *Upward Monotone*.

Domains We construct sentences based on existing text from four domains: News (drawn from articles linked on Google News[7]), Reddit (from threads linked on the Front Page[8]), Wikipedia (from Featured Articles[9]), and academic papers drawn from the proceedings of recent ACL conferences. We include 100 sentence pairs constructed from each source, as well as 150 artificially-constructed sentence pairs.

Annotation Process We begin with an initial set of fine-grained semantic phenomena, using the

[7] `news.google.com`
[8] `reddit.com`
[9] `en.wikipedia.org/wiki/Wikipedia:Featured_articles`

Coarse-Grained Categories	Fine-Grained Categories
Lexical Semantics	Lexical Entailment, Morphological Negation, Factivity, Symmetry/Collectivity, Redundancy, Named Entities, Quantifiers
Predicate-Argument Structure	Core Arguments, Prepositional Phrases, Ellipsis/Implicits, Anaphora/Coreference Active/Passive, Nominalization, Genitives/Partitives, Datives, Relative Clauses, Coordination Scope, Intersectivity, Restrictivity
Logic	Negation, Double Negation, Intervals/Numbers, Conjunction, Disjunction, Conditionals, Universal, Existential, Temporal, Upward Monotone, Downward Monotone, Non-Monotone
Knowledge	Common Sense, World Knowledge

Table 3: The types of linguistic phenomena annotated in the diagnostic dataset, organized under four major categories.

categories in the FraCaS test suite (Cooper et al., 1996) as a starting point, while also generalizing to include lexical semantics, common sense, and world knowledge. We gather examples by searching through text in each domain and locating example sentences that can be easily modified to involve one of the chosen phenomena (or that involves one already). We then modify the sentence further to produce the other sentence in an NLI pair. In many cases, we make these modifications small, in order to encourage high lexical and structural overlap among the sentence pairs—which may make the examples more difficult for models that rely on lexical overlap as an indicator for entailment. We then label the NLI relations between the sentences in both directions (considering each sentence alternatively as the premise), producing two labeled examples for each pair. Where possible, we produce several pairs with different labels for a single sentence, to have minimal sets of sentence pairs that are lexically and structurally very similar but correspond to different entailment relationships. After finalizing the categories, we gathered a minimum number of examples in each fine-grained category from each domain to ensure a baseline level of diversity.

In total, we gather 550 sentence pairs, for 1100 entailment examples. The labels are 42% entailment, 35% neutral, and 23% contradiction.

Auditing In light of recent work showing that crowdsourced data often contains artifacts which can be exploited to perform well without solving the intended task (Schwartz et al., 2017; Gururangan et al., 2018), we perform an audit of our manually curated data as a sanity check. We reproduce the methodology of Gururangan et al. (2018), training fastText classifiers (Joulin et al., 2016) to predict entailment labels on SNLI and MultiNLI using only the hypothesis as input. Testing these on the diagnostic data, accuracies are 32.7% and 36.4%—very close to chance—showing that the data does not suffer from artifacts of this specific kind. We also evaluate state-of-the-art NLI models on the diagnostic dataset and find their overall performance to be rather weak, further suggesting that no easily-gameable artifacts present in existing training data are abundant in the diagnostic dataset (see Section 6).

Evaluation Since the class distribution in the diagnostic set is not uniform (and is even less so within each category), we propose using R_3, a three-class generalization of the Matthews correlation coefficient, as the evaluation metric. This coefficient was introduced by Gorodkin (2004) as R_K, a generalization of the Pearson correlation that works for K dimensions by averaging the square error from the mean value in each dimension, i.e., calculating the full covariance between the input and output. In the discrete case, it generalizes Matthews correlation, where a value of 1 means perfect correlation and 0 means random chance.

Intended Use Because these analysis examples are hand-picked to address certain phenomena, we expect that they will not be representative of the distribution of language as a whole, even in the targeted domains. However, NLI is a task with no natural input distribution. We deliberately select sentences that we hope will be able to provide insight into what models are doing, what phenomena they catch on to, and where are they limited. This means that the raw performance numbers on the analysis set should be taken with a grain of salt. The set is provided not as a benchmark, but as an analysis tool to paint in broad strokes the kinds

Tags	Premise	Hypothesis	Fwd	Bwd
UQuant	Our deepest sympathies are with all those affected by this accident.	Our deepest sympathies are with a victim who was affected by this accident.	E	N
MNeg	We built our society on unclean energy.	We built our society on clean energy.	C	C
MNeg, 2Neg	The market is about to get harder, but not impossible to navigate.	The market is about to get harder, but possible to navigate.	E	E
2Neg	I have never seen a hummingbird not flying.	I have always seen hummingbirds flying.	E	E
2Neg, Coref	It's not the case that there is no rabbi at this wedding; he is right there standing behind that tree.	A rabbi is at this wedding, standing right there standing behind that tree.	E	E

Table 4: Examples from the diagnostic evaluation. Tags are *Universal Quantification* (UQuant), *Morphological Negation* (MNeg), *Double Negation* (2Neg), and *Anaphora/Coreference* (Coref). Other tags on these examples are omitted for brevity.

of phenomena a model may or may not capture, and to provide a set of examples that can serve for error analysis, qualitative model comparison, and development of adversarial examples that expose a model's weaknesses.

5 Baselines

As baselines, we provide performance numbers for a relatively simple multi-task learning model trained from scratch on the benchmark tasks, as well as several more sophisticated variants that utilize recent developments in transfer learning. We also evaluate a sample of competitive existing sentence representation models, where we only train task-specific classifiers on top of the representations they produce.

5.1 Multi-task Architecture

Our simplest baseline is based on sentence-to-vector encoders, and sets aside GLUE's ability to evaluate models with more complex structures. Taking inspiration from Conneau et al. (2017), the model uses a BiLSTM with temporal max-pooling and 300-dimensional GloVe word embeddings (Pennington et al., 2014) trained on 840B Common Crawl. For single-sentence tasks, we process the sentence and pass the resulting vector to a classifier. For sentence-pair tasks, we process sentences independently to produce vectors u, v, and pass $[u; v; |u - v|; u * v]$ to a classifier. We experiment with logistic regression and a multi-layer perceptron with a single hidden layer for classifiers, leaving the choice as a hyperparameter to tune.

For sentence-pair tasks, we take advantage of GLUE's indifference to model architecture by incorporating a matrix attention mechanism between the two sentences. By explicitly modeling the interaction between sentences, our model is strictly outside of the sentence-to-vector paradigm. We follow standard practice to contextualize each token with attention. Given two sequences of hidden states $u_1, u_2, \ldots, u_M$ and $v_1, v_2, \ldots, v_N$, the attention mechanism is implemented by first computing a matrix H where $H_{ij} = u_i \cdot v_j$. For each u_i, we get attention weights α_i by taking a softmax over the i^{th} row of H, and get the corresponding context vector $\tilde{v}_i = \sum_j \alpha_{ij} v_j$ by taking the attention-weighted sum of the v_j. We pass a second BiLSTM with max pooling over the sequence $[u_1; v_1], \ldots [u_M; v_M]$ to produce u'. We process the v_j vectors in a symmetric manner to obtain v'. Finally, we feed $[u'; v'; |u' - v'|; u' * v']$ into a classifier for each task.

Incorporating Transfer Learning We also augment our base non-attentive model with two recently proposed methods for transfer learning in NLP: ELMo (Peters et al., 2018) and CoVe (McCann et al., 2017). Both use pretrained models that produce contextual word embeddings via some transformation of the underlying model's hidden states.

ELMo uses a pair of two-layer neural language models (one forward, one backward) trained on the One Billion Word Benchmark (Chelba et al., 2013). A word's contextual embedding is produced by taking a linear combination of the corresponding hidden states on each layer. We follow the authors' recommendations[10] and use the ELMo embeddings in place of any other embed-

[10]github.com/allenai/allennlp/blob/master/tutorials/how_to/elmo.md

Model	Avg	Single Sent		Similarity and Paraphrase			Natural Language Inference			
		CoLA	SST-2	MRPC	QQP	STS-B	MNLI	QNLI	RTE	WNLI
Single-Task Training										
BiLSTM +ELMo	**60.3**	**22.2**	**89.3**	70.3/79.6	84.8/64.2	72.3/70.9	**77.1/76.6**$^{\mathcal{A}}$	**82.3**$^{\mathcal{A}}$	48.3$^{\mathcal{A}}$	63
Multi-Task Training										
BiLSTM	55.3	0.0	84.9	72.6/80.9	85.5/63.4	71.8/70.1	69.3/69.1	75.6	57.6	43
+Attn	56.3	0.0	83.8	74.1/82.1$^{\mathcal{A}}$	85.1/**63.6**$^{\mathcal{A}}$	71.1/69.8$^{\mathcal{A}}$	72.4/72.2$^{\mathcal{A}}$	78.5$^{\mathcal{A}}$	**61.1**$^{\mathcal{A}}$	44$^{\mathcal{A}}$
+ELMo	55.8	0.9	88.4	72.7/82.6	79.0/58.9	73.3/72.0	71.3/71.8	75.8	56.0	46
+CoVe	56.7	1.8	85.0	73.5/81.4	85.2/63.5	73.4/72.1	70.5/70.5	75.6	57.6	52
Pre-trained Sentence Representation Models										
CBoW	52.2	0.0	80.1	72.6/81.1	79.4/51.2	61.2/59.0	55.7/56.3	71.5	54.8	63
Skip-Thought	55.4	0.0	82.1	73.4/82.4	81.7/56.1	71.7/69.8	62.8/62.7	72.6	55.1	64
InferSent	58.1	2.8	84.8	75.7/82.8	**86.1**/62.7	75.8/75.7	66.0/65.6	73.7	59.3	**65**
DisSent	56.4	9.9	83.9	76.7/83.7	85.3/62.7	66.2/65.1	57.7/58.0	68.0	59.5	**65**
GenSen	59.2	1.4	83.6	**78.2**/**84.6**	83.2/59.8	**79.0**/**79.4**	71.2/71.0	78.7	59.9	**65**

Table 5: Performances on the benchmark tasks for different models. Bold denotes best results per task overall; underline denotes best results per task within a section; $\mathcal{A}$ denotes models using attention. For MNLI, we report accuracy on the matched / mismatched test splits. For MRPC and Quora, we report accuracy / F1. For STS-B, we report Pearson / Spearman correlation, scaled to be in [-100, 100]. For CoLA, we report Matthews correlation, scaled to be in [-100, 100]. For all other tasks we report accuracy (%). We compute a macro-average score in the style of SentEval by taking the average across all tasks, first averaging the metrics within each tasks for tasks with more than one reported metric.

dings.

CoVe uses a sequence-to-sequence model with a two-layer BiLSTM encoder trained for English-to-German translation. The CoVe vector $C(w_i)$ of a word is the corresponding hidden state of the top-layer LSTM. As per the original work, we concatenate the CoVe vectors to the GloVe word embeddings.

5.2 Multi-task Training

These four models (BiLSTM, BiLSTM +Attn, BiLSTM +ELMo, BiLSTM +CoVe) are jointly trained on all tasks, with the primary BiLSTM encoder shared between all task-specific classifiers. To perform multi-task training, we randomly pick an ordering on the tasks and train on 10% of a task's training data for each task in that order. We repeat this process 10 times between validation checks, so that we roughly train on all training examples for each task once between checks. We use the previously defined macro-average as the validation metric, where for tasks without predetermined development sets, we reserve 10% of the training data for validation.

We train our models with stochastic gradient descent using batch size 128, and multiply the learning rate by .2 whenever validation performance does not improve. We stop training when the learning rate drops below 10^{-5} or validation performance does not improve after 5 evaluations. We tune hyperparameters with random search over 30 runs on macro-average development set performance. Our best model is a two layer BiLSTM that is 1500-dimensional per direction. We evaluate our all our BiLSTM-based models with these settings.

5.3 Single-task Training

We use the same training procedure to train an instance of the model with ELMo on each task separately. For tuning hyperparameters per task, we use random search on that task's metrics evaluated on the development set. We tune the same hyperparameters as in the multi-task setting, except we also tune whether or not to use attention (for pair tasks only), and whether to use SGD or Adam (Kingma and Ba, 2014).

5.4 Sentence Representation Models

Finally, we evaluate a number of established sentence-to-vector encoder models using our suite. Specifically, we investigate:

1. CBoW: the average of the GloVe embeddings of the tokens in the sentence.

2. Skip-Thought (Kiros et al., 2015): a sequence-to-sequence(s) model trained to generate the previous and next sentences given the middle sentence. After training, the

model's encoder is taken as a sentence encoder. We use the original pretrained model[11] trained on sequences of sentences from the Toronto Book Corpus (Zhu et al. 2015, TBC).

3. InferSent (Conneau et al., 2017): a BiLSTM with max-pooling trained on MNLI and SNLI.

4. DisSent (Nie et al., 2017): a BiLSTM with max-pooling trained to predict the discourse marker (e.g. "because", "so", etc.) relating two sentences on data derived from TBC (Zhu et al., 2015). We use the variant trained to predict eight discourse marker types.

5. GenSen (Subramanian et al., 2018): a sequence-to-sequence model trained on a variety of supervised and unsupervised objectives. We use a variant of the model trained on both MNLI and SNLI, the Skip-Thought objective on TBC, and a constituency parsing objective on the One Billion Word Benchmark.

We use pretrained versions of these models, fix their parameters, learn task-specific classifiers on top of the sentence representations that they produce. We use the SentEval framework to train the classifiers.

6 Benchmark Results

We present performance on the main benchmark in Table 5. For multi-task models, we average performance over five runs; for single-task models, we use only one run.

We find that the single-task baselines have the best performance among all models on SST-2, MNLI, and QNLI, while the lagging behind multi-task trained models on MRPC, STS-B, and RTE. For MRPC and RTE in particular, the single-task baselines are close to majority class baselines, indicating the inherent difficulty of these tasks and the potential of transfer learning approaches. On QQP, the best multi-task trained models slightly outperform the single-task baseline.

For multi-task trained baselines, we find that almost no model does significantly better on CoLA or WNLI than performance from predicting majority class (0.0 and 63, respectively), which highlights the difficulty of current models to generalize to these tasks. The notable exception is DisSent, which does better than other multi-task models on CoLA. A possible explanation is that DisSent is trained using a discourse-based objective, which might be more sensitive to grammaticality. However, DisSent underperforms other multi-task models on more data-rich tasks such as MNLI and QNLI. This result demonstrates the utility of GLUE: by assembling a wide variety of tasks, it highlights the relative strengths and weaknesses of various models.

Among our multi-task BiLSTM models, using attention yields a noticeable improvement over the vanilla BiLSTM for all tasks involving sentence pairs. When using ELMo or CoVe, we see improvements for nearly all tasks. There is also a performance gap between all variants of our multi-task BiLSTM model and the best models that use pre-trained sentence representations (GenSen and Infersent), demonstrating the utility of transfer via pre-training on an auxiliary task.

Among the pretrained sentence representation models, we observe relatively consistent per-task and aggregate performance gains moving from CBoW to Skip-Thought to DisSent, to Infersent and GenSen. The latter two show competitive performance on various tasks, with GenSen slightly edging out InferSent in aggregate.

Model	LS	PAS	L	K	All
BiLSTM	13	27	15	15	20
BiLSTM +Attn$^{\mathcal{A}}$	26	**32**	**24**	**18**	**27**
BiLSTM +ELMo	14	22	14	17	17
BiLSTM +CoVe	17	30	17	14	21
CBoW	09	13	08	10	10
Skip-Thought	02	25	09	08	12
InferSent	17	18	15	13	19
DisSent	09	14	11	15	13
GenSen	**27**	27	14	10	20

Table 6: Results on the diagnostic set. $\mathcal{A}$ denotes models using attention. All numbers in the table are R_3 coefficients between gold and predicted labels within each category (percentage). The categories are *Lexical Semantics* (**L**), *Predicate-Argument Structure* (**PAS**), *Logic* (**L**), and *Knowledge and Common Sense* (**K**).

7 Analysis

By running all of the models on the diagnostic set, we get a breakdown of their performance across a set of modeling-relevant phenomena. Overall re-

[11] github.com/ryankiros/skip-thoughts

Gold \Prediction	All	E	C	N
All		65	16	19
E	42	34	3	4
C	23	11	8	4
N	35	19	5	11

(a) Confusion matrix for BiLSTM +Attn (percentages).

Model	E	C	N
BiLSTM	71	16	13
BiLSTM +Attn$^{\mathcal{A}}$	**65**	16	**19**
BiLSTM +ELMo	81	9	10
BiLSTM +Cove	75	13	13
CBoW	84	7	9
SkipThought	80	8	12
InferSent	68	**21**	11
DisSent	73	18	8
GenSen	74	15	11
Gold	42	23	35

(b) Output class distributions (percentages). Bolded numbers are closest to the gold distribution.

Figure 1: Partial output of GLUE's error analysis, aggregated across our models.

Model	UQuant	MNeg	2Neg	Coref
BiLSTM	67	13	5	24
BiLSTM +Attn$^{\mathcal{A}}$	**85**	**64**	11	20
BiLSTM +ELMo	77	60	-8	18
BiLSTM +CoVe	71	34	**28**	**39**
CBoW	16	0	13	21
SkipThought	61	6	-2	30
InferSent	64	51	-22	26
DisSent	70	34	-20	21
GenSen	78	**64**	5	26

Table 7: Model performance in terms of R_3 (scaled by 100) on selected fine-grained categories for analysis. The categories are *Universal Quantification* (**UQuant**), *Morphological Negation* (**MNeg**), *Double Negation* (**2Neg**), and *Anaphora/Coreference* (**Coref**).

sults are presented in Table 6.

Overall Performance Performance is very low across the board: the highest total score (27) still denotes poor absolute performance. Scores on the Predicate-Argument Structure category tend to be higher across all models, while Knowledge category scores are lower. However, these trends do not necessarily reflect that our models understand sentence structure better than world knowledge or common sense; these numbers are not directly comparable. Rather, numbers should be compared between models within each category.

One notable trend is the high performance of the BiLSTM +Attn model: though it does not outperform most of the pretrained sentence representation methods (InferSent, DisSent, GenSen) on GLUE's main benchmark tasks, it performs best or competitively on all categories of the diagnostic set.

Domain Shift & Class Priors GLUE's online platform also provides a submitted model's predicted class distributions and confusion matrices. We provide an example in Figure 1. One point is immediately clear: all models severely underpredict *neutral* and over-predict *entailment*. This is perhaps indicative of the models' inability to generalize and adapt to new domains. We hypothesize that they learned to treat high lexical overlap as a strong sign of entailment, and that surgical addition of new information to the hypothesis (as in the case of neutral instances in the diagnostic set) might go unnoticed. Indeed, the attention-based model seems more sensitive to the neutral class, and is perhaps better at detecting small sets of unaligned tokens because it explicitly tries to model these alignments.

Linguistic Phenomena While performance metrics on the coarse-grained categories give us broad strokes that we can use to compare models, we can gain a better understanding of the models' capabilities by drilling down into the fine-grained subcategories. The GLUE platform reports scores for every fine-grained category; we present here a few highlights in Table 7. To help interpret these results, we list some examples from each fine-grained category, along with model predictions, in Table 4.

The Universal Quantification category appears easy for most of the models; looking at examples, it seems that when universal quantification as a phenomenon is isolated, catching on to lexical cues such as *all* often suffices to solve our examples. Morphological negation examples are superficially similar, but the systems find it more difficult. On the other hand, *double* negation appears to be adversarially difficult for models to recognize, with the exception of BiLSTM +CoVe; this is perhaps due to the translation signal, which can match phrases like "not bad" and "okay" to the same expression in a foreign language. A similar advantage, though less acute, appears when using CoVe on coreference examples.

Overall, there is some evidence that going beyond sentence-to-vector representations might aid performance on out-of-domain data (as with BiLSTM +Attn) and that representations like ELMo and CoVe encode important linguistic information that is specific to their supervision signal. Our platform and diagnostic dataset should support future inquiries into these issues, so we can better understand our models' generalization behavior and what kind of information they encode.

8 Conclusion

We introduce GLUE, a platform and collection of resources for training, evaluating, and analyzing general natural language understanding systems. When evaluating existing models on the main GLUE benchmark, we find that none are able to substantially outperform a relatively simple baseline of training a separate model for each constituent task. When evaluating these models on our diagnostic dataset, we find that they spectacularly fail on a wide range of linguistic phenomena. The question of how to design general-purpose NLU models thus remains unanswered. We believe that GLUE, and the generality it promotes, can provide fertile soil for addressing this open challenge.

Acknowledgments

We thank Ellie Pavlick, Tal Linzen, Kyunghyun Cho, and Nikita Nangia for their comments on this work at its early stages, and we thank Ernie Davis, Alex Warstadt, and Quora's Nikhil Dandekar and Kornel Csernai for providing access to private evaluation data. This project has benefited from financial support to SB by Google, Tencent Holdings, and Samsung Research.

References

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

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, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC*.

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

Daniel Cer, 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*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *2nd Workshop on Evaluating Vector Space Representations for NLP*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *LREC 2018*.

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.

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

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

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

T. Mark Ellison, editor. 1997. *Computational Natural Language Learning: Proceedings of the 1997 Meeting of the ACL Special Interest Group in Natural Language Learning*. Association for Computational Linguistics.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *First Workshop on Building Linguistically Generalizable NLP Systems*.

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.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *Proceedings of ICLR 2018*.

J. Gorodkin. 2004. Comparing two k-category assignments by a k-category correlation coefficient. *Comput. Biol. Chem.*, 28(5-6):367–374.

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*.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. In *Proceedings of EMNLP 2017*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *Proceedings of NAACL 2016*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *ICLR 2015*.

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.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

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.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. *CoRR*, abs/1705.06476.

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

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

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.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2017. Sluice networks: Learning what to share between loosely related tasks. *arXiv preprint arXiv:1705.08142*.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79.

Carson T Schütze. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proc. of CoNLL*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *ICLR 2017*.

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.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 231–235.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. In *ICLR 2016*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 996–1005.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

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

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. In *ICLR 2017*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.