GLUE and SuperGLUE Benchmark

By: Hiva Mohammadzadeh

General Language Understanding Evaluation(GLUE) Benchmark

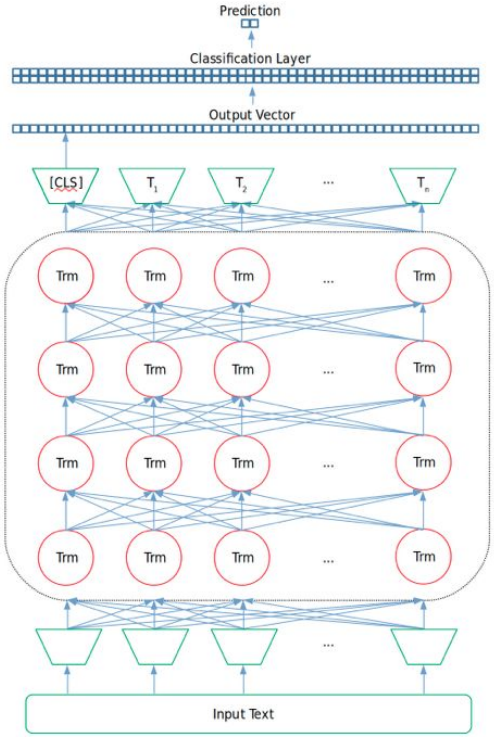
A collection of resources for training, evaluating, and analyzing natural language understanding systems.

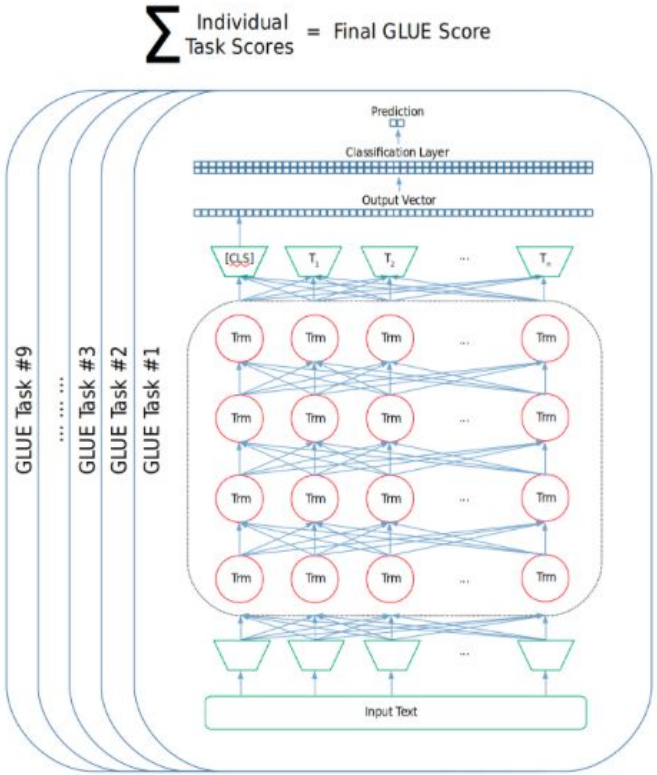
GLUE Benchmark

- Collection of tools for evaluating the performance of models which includes:
  - Nine language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty.
  - A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language.
  - A public leaderboard for tracking performance on the benchmark and a dashboard for visualizing the performance of models on the diagnostic set.
  - A handcrafted diagnostic test suite that enables detailed linguistic analysis of models.
  - All tasks are single sentence or sentence pair classification except STS-B which is a regression task.
- The tasks include question answering, sentiment analysis, and textual entailment, and an associated online platform for model evaluation, comparison, and analysis.

GLUE Score Calculation

- It lets researchers compare, in a single number, models against one another.
- To test the model, remove the pretraining classification layer and replace it with one that accommodates the output of the GLUE task.
- Then train and then score the model on all nine tasks, and the resulting average score of those nine tasks is the model's final performance score.





Tasks

- **CoLA (Corpus of Linguistic Acceptability)** - determines if a given sentence is grammatically correct. (Single-Sentence Task)
- **SST-2 (Stanford Sentiment Treebank)** - determines the sentiment of a given sentence (positive or negative). (Single-Sentence Task)
- **MRPC (Microsoft Research Paraphrase Corpus)** - determines if two sentences are semantically equivalent. (Similarity and Paraphrase Task)
- **QQP (Quora Question Pairs)** - determines if two questions are semantically equivalent. (Similarity and Paraphrase Task)
- **STS-B (Semantic Textual Similarity Benchmark)** - determines the degree of semantic similarity between two sentences. (Similarity and Paraphrase Task)
- **MNLI (Multi-Genre Natural Language Inference**) - determines the relationship between two given sentences (entailment, neutral or contradiction). (Inference Task)
- **QNLI (Question Natural Language Inference)** - determines if a given question can be answered based on a given context (entailment or not). (Inference Task)
- **RTE (Recognizing Textual Entailment)** - determines if a given hypothesis can be inferred from a given text (entailment or not). (Inference Task)
- **WNLI (Winograd Schema Challenge)** - determines the relationship between two given sentences based on a pronoun resolution problem. (Inference Task)

Corpus of Linguistic Acceptability (CoLA)

- Single-Sentence Task
- It is used to determine if a given sentence is grammatically correct (binary Classification Task).
- It has 10657 sentences from 23 linguistics publications which includes 8551 sentences for training, 1043 sentences for validation, 1063 sentences for test set.
- Size of downloaded dataset files: 0.38 MB, Size of the generated dataset: 0.61 MB, Total amount of disk used: 0.99 MB
- The authors use Matthews Correlation Coefficient (MCC) as the evaluation metric to evaluate the performance on unbalanced binary classification.
- Example from the dataset:
  - { "sentence": "Our friends won't buy this analysis, let alone the next one we propose.", "label": 1, "id": 0 }

Stanford Sentiment Treebank (SST-2)

- Single-Sentence Task
- It is used to determines the sentiment of a given sentence (positive or negative) (Binary Classification Task).
- It has 70043 sentences which includes 67349 sentences for training, 872 sentences for validation, 1822 sentences for test set.
- The sentences are taken from movie reviews.
- The authors just use accuracy as the evaluation metric, measuring the percentage of correctly classified sentences in the dataset.
- Example:
  - {"sentence": "that loves its characters and communicates something rather beautiful about human nature “, "label": 1}

Microsoft Research Paraphrase Corpus (MRPC)

- Similarity and Paraphrase Task
- It is used to determine if two sentences are semantically equivalent (Binary Classification Task).
- It has 5800 sentences, which includes 3.7k sentences for training, and 1.7k sentences for test set.
- The sentences are taken from online news sources.
- The classes are imbalanced (68% positive), and therefore the authors use both accuracy and F1 score.
- Example:
  - {“Sentence1”: “The world's two largest automakers said their U.S. sales declined more than predicted last month as a late summer sales frenzy caused more of an industry backlash than expected.”, “Sentence2”: “Domestic sales at both GM and No. 2 Ford Motor Co. declined more than predicted as a late summer sales frenzy prompted a larger-than-expected industry backlash.”, label: 1}

Quora Question Pairs (QQP)

- Similarity and Paraphrase Task
- It is used to determine if two questions are semantically equivalent (Binary Classification Task).
- It has 795242 sentences which includes 363846 sentences for training, 40430 sentences for validation, 390966 sentences for test set.
- The sentences are taken from the community question-answering website Quora.
- The classes are imbalanced (63% negative), and therefore the authors use both accuracy and F1 score.
- Example:
  - { “question1”: “How is the life of a math student? Could you describe your own experiences?”, “question2”: “Which level of prepration is enough for the exam jlpt5?”, “Is_duplicate”: 0}

Semantic Textual Similarity Benchmark (STS-B)

- Similarity and Paraphrase Task
- It is used to determine the degree of semantic similarity using a similarity score from 1 to 5 between two sentences.
- It has 8302 sentences which includes 5712 sentences for training, 1470 sentences for validation, 1120 sentences for test set.
- The sentences are taken from news headlines, video and image captions, and natural language inference data.
- The authors evaluate the model using Pearson (evaluates the linear relationship between two continuous variables) and Spearman (evaluates the monotonic relationship between two variables) Correlation coefficients.
- Example:
  - { “sentence1”: “A plane is taking off”, “sentence2”: “An air plane is taking off”, “Score”: 5.000}

Multi-Genre Natural Language Inference (MNLI)

- Inference Task
- It is used to determines the relationship between two given sentences, premise and hypothesis sentences (3 classes: entailment (0), neutral (1) or contradiction(2)).
- It has 10657 sentences which includes 392702 sentences for training, 9815 sentences for validation_matched, 9832 sentences for validation_mismatched, 9796 sentences for test_matched, and 9847 for test_mismatched.
- The premise sentences are gathered from ten different sources, including transcribed speech, fiction, and government reports.
- Size of downloaded dataset files: 312.78 MB, Size of the generated dataset: 82.47 MB, Total amount of disk used: 395.26 MB.
- The authors evaluate on both the matched and mismatched sections.
- Example:
  - { "premise": "Conceptually cream skimming has two basic dimensions - product and geography.", "hypothesis": "Product and geography are what make cream skimming work.", "label": 1, "idx": 0 }

Question Natural Language Inference (QNLI)

- Inference Task
- It is used to determine if a given question can be answered based on a given context (entailment or not).
- It has 110k sentences, which includes 105k sentences for training, and 5.4k sentences for test set.
- The sentences are taken from the Stanford Question Answering Dataset and the Wikipedia.
- Example:
  - { “question”: “Which missile batteries often have individual launchers several kilometres from one another?”, “sentence”: “When MANPADS is operated by specialists, batteries may have several dozen teams deploying separately in small sections; self-propelled air defence guns may deploy in pairs.”, “label”: “not_entailment”}

Recognizing Textual Entailment (RTE)

- Inference Task
- It is used to determine if a given hypothesis can be inferred from a given text (entailment or not).
- It has 5.5k sentences, which includes 2.5k sentences for training, and 3k sentences for test set.
- The sentences are taken from a series of annual textual entailment challenges, news and Wikipedia.
- Example:
  - { “sentence1”: “No Weapons of Mass Destruction Found in Iraq Yet.” , “sentence2”: “Weapons of Mass Destruction Found in Iraq.” , “label”: “not_entailment”}

Winograd Schema Challenge (WNLI)

- Inference Task
- It is used to determine the relationship between two given sentences based on a pronoun resolution problem. System must read a sentence with a pronoun and select the referent of that pronoun from a list of choices.
- It has 1000 sentences, which includes 634 sentences for training, and 146 sentences for test set.
- The sentences are taken from fiction books.
- The included training set is balanced but the test set is imbalanced (65% not entailment).
- Each example is evaluated separately, so there is not a systematic correspondence between a model's score on this task and its score on the unconverted original task.
- Example:
  - { "sentence1": "I stuck a pin through a carrot. When I pulled the pin out, it had a hole." , "sentence2": "The carrot had a hole." , "label": 1}

SuperGLUE Benchmark

A new benchmark upgraded from GLUE with a new set of more difficult language understanding tasks, a software toolkit, and a public leaderboard.

Improvements

- **More challenging tasks:** SuperGLUE retains the two hardest tasks in GLUE identified from those submitted to an open call for task proposals.
- **More diverse task formats:** The task formats in GLUE are limited to sentence and sentence pair classification. We expand the set of task formats in SuperGLUE to include coreference resolution and question answering (QA).
- **Comprehensive human baselines:** We include human performance estimates for all benchmark tasks, which verify that substantial headroom exists between a strong BERT-based baseline and human performance.
- **Improved code support:** SuperGLUE is distributed with a new, modular toolkit for work on pretraining, multi-task learning, and transfer learning in NLP, built around standard tools including PyTorch (Paszke et al., 2017) and AllenNLP (Gardner et al., 2017).
- **Refined usage rules:** The conditions for inclusion on the SuperGLUE leaderboard have been revamped to ensure fair competition, an informative leaderboard, and full credit assignment to data and task creators.

Tasks

1. **BoolQ (Boolean Questions)** - determines whether a given sentence answers a yes/no question.
2. **CB (CommitmentBank)** - determines the degree of commitment expressed in a given sentence.
3. **COPA (Choice of Plausible Alternatives)** - determines which of two alternatives is more plausible as the cause or effect of a given event.
4. **MultiRC (Multi-Sentence Reading Comprehension)** - determines the answer to a question based on multiple passages of text.
5. **ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset)** - determines the answer to a question based on a given passage of text and commonsense knowledge.
6. **RTE (Recognizing Textual Entailment)** - determines if a given hypothesis can be inferred from a given text (entailment or not).
7. **WiC (Word-in-Context**) - determines if a word is used with the same sense in two different sentences.
8. **WSC (Winograd Schema Challenge)** - determines the referent of a pronoun in a sentence based on commonsense knowledge

Boolean Questions (BoolQ)

- Question Answering Task
- It is used to determine whether a given sentence answers a yes/no question.
- It has 15942 sentences which includes 9427 sentences for training, 3270 sentences for validation, 3245 sentences for test set.
- The questions are taken from queries of google search engine and they are paired with a paragraph from Wikipedia.
- The authors use accuracy as the evaluation metric to evaluate the performance of the task.
- Example:
  - { "question": "calcium carbide cac2 is the raw material for the production of acetylene", "passage": "Calcium carbide -- Calcium carbide is a chemical compound with the chemical formula of CaC. Its main use industrially is in the production of acetylene and calcium cyanamide.", "idx": 13, "label": true}

Commitment Bank (CB)

- Natural Language inference Task
- It is used to determine the degree of commitment expressed in a given sentence.
- It has 557 sentences which includes 250 sentences for training, 57 sentences for validation, 250 sentences for test set.
- The sentences are taken from various sources.
- The authors use accuracy and the F1 score as the evaluation metric to evaluate the performance of the task. For multi-class F1, they compute the unweighted average of the F1 per class.
- Example:
  - {“premise": "It was a complex language. Not written down but handed down. One might say it was peeled down.", "hypothesis": "the language was peeled down", "label": "entailment", "idx": 0}

Choice of Plausible Alternatives (COPA)

- Question Answering Task
- It is used to determine which of two alternatives is more plausible as the cause or effect of a given event.
- It has 1000 sentences which includes 400 sentences for training, 100 sentences for validation, 500 sentences for test set.
- The sentences are taken from blogs and photography-related encyclopedia.
- The authors use accuracy as the evaluation metric to evaluate the performance of the task.
- Example:
  - { "premise": "The man got a discount on his groceries.", "choice1": "He greeted the cashier.", "choice2": "He used a coupon.", "question": "cause", "label": 1, "idx": 7}

Multi-Sentence Reading Comprehension (MultiRC)

- Question Answering Task
- It is used to determine the answer to a question based on multiple passages of text. The system predicts which answers are true or false.
- It has 7853 sentences which includes 5100 sentences for training, 953 sentences for validation, 1800 sentences for test set.
- The sentences are taken from various sources including news, fiction, and historical text.
- The authors use binary F1 score on all answer-option and exact match (EM) of each question's set of answers to evaluate the performance of the task.
- This task is used because some questions have multiple answers, some questions require drawing facts from a passage, and the API matches.
- Example:

MultiRC

Paragraph: *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*
Question: *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (T), *No* (F), *Yes* (T), *No, she didn't recover* (F), *Yes, she was at Susan's party* (T)

Reading Comprehension with Commonsense Reasoning Dataset (ReCoRD)

- Question Answering Task (Multiple choice QA task)
- It is used to determine the answer to a question based on a given passage of text and commonsense knowledge.
- It has 121k sentences which includes 101k sentences for training, 10k sentences for validation, 10k sentences for test set.
- The sentences are taken from news articles such as CNN and Daily Mail.
- The authors use max token-level F1 score and exact match (EM) as the evaluation metric to evaluate the performance of the task.
- Example:

ReCoRD

Paragraph: *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electorcal Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood*

Query For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency **Correct Entities:** US

Recognizing Textual Entailment (RTE)

- Natural Language inference Task
- It is used to determine if a given hypothesis can be inferred from a given text (entailment or not_entailment).
- It has 3078 sentences which includes 2500 sentences for training, 278 sentences for validation, 300 sentences for test set.
- The sentences are taken from news articles and the Wikipedia.
- The authors use accuracy as the evaluation metric to evaluate the performance of the task.
- One of the tasks that benefits from transfer learning from ~56% to 86.3% accuracy.
- Example:
  - { "premise": "Oil prices fall back as Yukos oil threat lifted", "hypothesis": "Oil prices rise.", "label": "not_entailment", "idx": 13}

Word-in-Context (WiC)

- Word Sense Disambiguation Task
- It is used to determine if a word is used with the same sense in two different sentences.
- It has 8038 sentences which includes 6000 sentences for training, 638 sentences for validation, 1400 sentences for test set.
- The sentences are taken from the WordNet, VerbNet and Wiktionary.
- The authors use accuracy as the evaluation metric to evaluate the performance of the task.
- Example:
  - {"word": "development", "sentence1": "The organism has reached a crucial stage in its development.", "sentence2": "Our news team brings you the latest developments.", "idx": 7, "label": false, "start1": 48, "start2": 36, "end1": 59, "end2": 48, "version": 1.1}

Winograd Schema Challenge (WSC)

- Coreference Resolution Task
- It is used to determine the referent of a pronoun in a sentence based on commonsense knowledge
- It has 804 sentences which includes 554 sentences for training, 104 sentences for validation, 146 sentences for test set.
- The sentences are taken from fiction books
- The authors use accuracy as the evaluation metric to evaluate the performance of the task.
- Requires everyday knowledge and commonsense reasoning to solve.
- Example:
  - { "text": "Sam Goodman 's biography of the Spartan general Xenophanes conveys a vivid sense of the difficulties he faced in his childhood.", "target": {"span2_index": 16, "span1_index": 1, "span1_text": "Goodman", "span2_text": "he"}, "idx": 6, "label": false}