The GLUE Benchmark

Vedant Palit

December 10, 2022

---

1 Introduction

About the Benchmark: The General Language Understanding Evaluation(GLUE) Benchmark is a collection of 9 sentence/ sentence-pair language understanding tasks built on established existing datasets. Datasets are chosen to cover a diverse-range of sizes, genres and degrees of difficulties.

What the Benchmark tests: Like Machine Learning, whose basic skillset involves linear algebra and calculus, Natural Language Processing involves two major skillsets for the agent -

- *Sentence Similarity:* Given a pair of sentences, the agent needs to identify any type of similarity in the sentences.
- *Entailment:* Given a sentence 1 to be true,implies another sentence 2 is also considered true. This is followed when it is said that sentence 1 entails sentence 2.

2 Tasks in GLUE Benchmark

There are 9 Tasks on the GLUE Benchmark Leaderboard:

- **The Corpus of Linguistic Acceptability(CoLA):** The dataset is a collection of English acceptability judgements drawn from books and journal articles on linguistic theory.
  - *Metric:* The Metric used to score is Matthew's Correlation Coefficient which ranges from -1 to 1, with 0 being the performance of uninformed guessing. The evaluation is based on **unbalanced binary classification**.
  - *Judgement:* The Task performs a binary classification of linguistic acceptability of an input sentence.
- **The Stanford Sentiment TreeBank(SST-2):** The dataset consists of single-sentence movie reviews from people followed by a binary annotation of the sentiment conveyed. A 1 conveys **Positive Sentiment** whereas a 0 conveys **Negative Sentiment**.

  - *Metric:* The Metric used is mainly the accuracy score of predictions made on newer movie reviews.
  - *Classification:* The sentiment values are between 0 to 1, in which a threshold is set(around 0.5), above which, the sentiment is positive **classified 1**, and below is negative **classified 0**.

- **The Microsoft Research Paraphrase Corpus:** The dataset consists of sentence pairs extracted from online news sources, with human annotations for semantic similarity of the sentences.
  - *Metric:* The Metric used is mainly the accuracy score of predictions. However due to an imbalance in the dataset, the F1 score is also used a metric.

- **The Quora Question Pairs:** The dataset contains question pairs from Quora, with annotations to identify semantic similarity of questions.
  - *Metric:* The Metric used is similar to **Microsoft Research Paraphrase Corpus** due to an imbalance in dataset.

- **The Semantic Textual Similarity Benchmark(STS-B):** The dataset consists of sentence pairs extracted from news headlines, video image captions and other NL inference data.
  - *Metric:* The metric to place the task on the leaderboard are Pearson and Spearman correlation coefficients. Each Human annotation is classified on a scale of 1-5 i.e A Multiclass Classification.

- **The Multi Genre Natural Language Inference Corpus(MLNI):** The dataset consists of crowd sourced sentence pairs, with entailment annotations.
  - *Entailment Annotation:* The sentence pair consists of a premise statement and a hypothesis sentence.The task is to check if premise entails hypothesis, contradicts it or is neutral to the hypothesis.
  - *Metric:* The Metric used is mainly the accuracy score of matchings.

- **The Stanford Question Answering Dataset(QLNI):** The dataset consists of a question-answer pair wherein, one sentence asks a question with another contextual sentence providing the answer to the question.
  - *Classification:* There are two possible results - **Positive and Negative**. A positive result shows that an answer can be extracted from the contextual sentence, and negation for negative.
  - *Metric:* The Metric used is mainly the accuracy score of matchings.

- **The Recognizing Textual Entailment:** The dataset comes from a series of annual textual entailment challenges. It is a binary classification of entailment wherein a 1 is entailing, a 0 is not entailing.

- **The Winograd Schema Challenge:** The task involves reading, wherein the system has to predict the referent of a pronoun from choices.

  - *Inner Workings:* The challenge involves building a sentence pair classification,with every possible referent replacing an ambigious pronoun. If the replaced sentence is entailed by the original sentence, it is classified as a **positive**.
  - *Metric:* The training set is perfectly balanced between entailment and not entailment, however due to an imbalance in the test set, the accuracy is used as a metric.

3 GLUE Benchmark and sufficiency in Natural Language Understanding

The GLUE Benchmark has a number of tasks, however there are a number of other benchmarks one of them being **decaNLP**. However, GLUE performs better than other benchmarks due to its varied set of tasks.

The **decaNLP benchmark** reduces all the evaluation tasks into a question-answering task which reduces the scope of the evaluation. Beyond this, these benchmarks also lack the error-analysis toolkits and leaderboards for appropriate evaluation of NLU tasks.

How is the GLUE Benchmark sufficient for NLU?

- *Scope of tasks:* The GLUE Benchmark focuses on utilising tasks that represent a larger scope of NLU- namely Question-Answering, Sentiment Analysis through sentence similarity and textual entailment.
- *Reliance on existing datasets:* The GLUE Benchmark uses already created widely existing datasets that have been accepted as interesting and challenging enough for NLU Tasks.
- *Less Imposed Restrictions:* The Benchmark does not impose restrictions on the existing model architecture of the tasks, other than restricting them to single sentence and pair extractions.

Inference: Hence the GLUE Benchmark is a sufficient and appropriate benchamrk for covering all the basic tasks related to Natural Language Understanding

---