



- Evaluating NLP systems on individual tasks seems a bit silly. A better approach to evaluating a model's language understanding capability is to evaluate it on many tasks. This is the idea behind the GLUE benchmark.

















The GLUE benchmark

2

5




















GUIEadair

## GLUE Tasks

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

# The GLUE benchmark

- Evaluating NLP systems on individual tasks seems a bit silly. A better approach to evaluating a model's language understanding capability is to evaluate it on many tasks. This is the idea behind the GLUE benchmark.

GLUE Tasks			
Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

GLUE Leaderboard



Georgetown  
University

# Lab 07: Project Kickoff