Experiment Protocol for CS224U, Spring 2020.

Abhishek Goswami
Microsoft / Redmond, WA
agoswami@microsoft.com

Abstract

This document sets the experiment protocol for CS224U, Spring 2020 quarter.

1 Hypothesis

A statement of the project's core hypothesis or hypotheses.

2 Data

A description of the dataset(s) that the project will use for evaluation.

3 Metrics

A description of the metrics that will form the basis for evaluation. We require at least one of these to be quantitative metrics, but we are very open-minded about which ones you choose. In requiring this, we are not saying that all work in NLU needs to be evaluated quantitatively, but rather just that we think it is a healthy requirement for our course.

4 Models

A description of the models that you'll be using as baselines, and a preliminary description of the model or models that will be the focus of your investigation. At this early stage, some aspects of these models might not yet be worked out, so preliminary descriptions are fine.

Categories	Train Set MNLI	Dev Set MNLI-(m/mm)	Test Set ANLI-r1
Contradiction	130,903	3,213 / 3,240	333
Entailment	130,899	3,479 / 3,463	334
Neutral	130,900	3,123 / 3,129	333

Table 1: Distribution for the Train/Dev/Test sets

5 General Reasoning

An explanation of how the data and models come together to inform your core hypothesis or hypotheses.

6 Progress Summary

what you have been done 2, what you still need to do, and any obstacles or concerns that might prevent your project from coming to fruition (Devlin et al., 2019)

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

System	Training Data Ratio							Dev Acc	Test Acc
	MNLI (400k)	MRPC (3k)	QQP (364k)	STSB (5k)	QNLI (105k)	RTE (2k)	WNLI (635)	MNLI-(m/mm) (10k / 10k)	ANLI-r1 (1k)
Random	1.0	-	-	-	-	-	-	33.8 / 33.2	33.3
Baseline	1.0	-	-	-	-	-	-	50.3 / 49.6	28.9
$\text{BERT}_{\text{BASE}}$	1.0	-	-	-	-	-	-	84.1 / 84.4	26.0
$\text{ROBERTa}_{\text{BASE}}$	1.0	-	-	-	-	-	-	87.7 / 87.6	28.5
$\text{ROBERTa}_{\text{LARGE}}$	1.0	-	-	-	-	-	-	92.0 / 90.0	46.0
$\text{BERT}_{\text{BASE}}$	0.1	0.1	0.1	-	-	-	-	76.9 / 77.2	23.2
with weak supervision	0.1	0.1	0.1	0.1	0.1	0.1	0.1	77.3 / 77.6	25.8
$\text{ROBERTa}_{\text{BASE}}$	0.1	0.1	0.1	-	-	-	-	83.6 / 84.1	26.0
with weak supervision	0.1	0.1	0.1	0.1	0.1	0.1	0.1	83.6 / 84.3	29.3

Table 2: Evaluation results on the Dev/Test sets