GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING

WHAT IS GLUE?

- TOOLS FOR EVALUATING AND ANALYZING THE PERFORMANCE OF MODELS ACROSS A DIVERSE SET OF EXISTING NLU TASKS.
  - 9 SENTENCE OR SENTENCE-PAIR NLU TASKS
  - AN ONLINE EVALUATION PLATFORM AND LEADERBOARD, BASED PRIMARILY ON PRIVATELY-HELD TEST DATA
  - AN EXPERT-CONSTRUCTED DIAGNOSTIC EVALUATION DATASET.
  - BASELINE RESULTS FOR SEVERAL MAJOR EXISTING APPROACHES TO SENTENCE REPRESENTATION LEARNING
- THE GOAL OF GLUE IS TO SPUR DEVELOPMENT OF GENERALIZABLE NLU SYSTEMS

SINGLE-SENTENCE TASKS

- COLA THE CORPUS OF LINGUISTIC ACCEPTABILITY – EACH SENTENCE IS ANNOTATED AS GRAMMATICAL OR UNGRAMMATICAL
  - TASK – DETERMINE GRAMMATICALITY OF SENTENCE
  - USE MATTHEW'S CORRELATION CO-EFFICIENT TO TEST

- SST-2 THE STANFORD SENTIMENT TREEBANK – SENTENCES EXTRACTED FROM MOVIE REVIEWS AND HUMAN ANNOTATIONS OF THEIR SENTIMENT.
  - TASK – DETERMINE SENTIMENT
  - POSITIVE/NEGATIVE CLASS SPLIT
  - SENTENCE-LEVEL LABELS.

SIMILARITY AND PARAPHRASE TASKS

- MRPC THE MICROSOFT RESEARCH PARAPHRASE CORPUS - SENTENCE PAIRS AUTOMATICALLY EXTRACTED FROM ONLINE NEWS SOURCES,
  - TASK – SEMANTIC EQUIVALENCE
  - 68% POSITIVE, 32% NEGATIVE
  - ACCURACY AND F1 SCORE.
- QQP THE QUORA QUESTION PAIRS - QUESTION PAIRS FROM QUORA
  - TASK – SEMANTIC EQUIVALENCE
  - 37% POSITIVE, 63% NEGATIVE
- STS-B THE SEMANTIC TEXTUAL SIMILARITY BENCHMARK – SENTENCE PAIRS
  - HUMAN-ANNOTATED WITH A SIMILARITY SCORE FROM 1 TO 5
  - TASK – PREDICT SIMILARITY SCORES
  - PEARSON AND SPEARMAN CORRELATION COEFFICIENTS

INFERENCE TASKS

- MNLI THE MULTI-GENRE NATURAL LANGUAGE INFERENCE CORPUS – CROWDSOURCED COLLECTION OF SENTENCE PAIRS WITH TEXTUAL ENTAILMENT ANNOTATIONS.
  - TASK – PREDICT ENTAILMENT, CONTRADICTION OR NEUTRAL
  - PRIVATE LABELS FROM THE AUTHORS
  - EVALUATE ON MATCHED (IN-DOMAIN) AND MISMATCHED (CROSS-DOMAIN) SECTIONS
- QNLI THE STANFORD QUESTION ANSWERING DATASET – QUESTION PARAGRAPH PAIRS
  - TASK - DETERMINE WHETHER THE CONTEXT SENTENCE CONTAINS THE ANSWER TO THE QUESTION

INFERENCE TASKS

- RTE THE RECOGNIZING TEXTUAL ENTAILMENT - ANNUAL CHALLENGES (RTE1, 2, 3, 5)ON TEXTUAL ENTAILMENT.
  - TWO-CLASS SPLIT
- WNLI THE WINOGRAD SCHEMA CHALLENGE - A READING COMPREHENSION TASK
  - READ A SENTENCE WITH A PRONOUN AND SELECT THE REFERENT OF THAT PRONOUN FROM A LIST OF CHOICES
  - TASK – PREDICT IF THE SENTENCE WITH THE PRONOUN SUBSTITUTED IS ENTAILED BY THE ORIGINAL SENTENCE

- EVALUATION
  - RUN THE SYSTEM ON THE PROVIDED TEST DATA FOR THE TASKS
  - UPLOAD THE RESULTS TO THE WEBSITE FOR SCORING
  - SHOWS PER-TASK SCORES, AS WELL AS A MACRO-AVERAGE OF THOSE SCORES TO DETERMINE A SYSTEM'S POSITION ON THE LEADERBOARD.
  - USE AN UNWEIGHTED AVERAGE OF THE METRICS AS THE SCORE FOR THE TASK WHEN COMPUTING THE OVERALL MACRO-AVERAGE.
- DATA AND BIAS
  - DON'T USE TRAINING DATA
- DIAGNOSTIC DATASET
  - MANUALLY-CURATED TEST SET (WITH PRIVATE LABELS) FOR THE ANALYSIS OF SYSTEM PERFORMANCE.

- DOMAINS
  - NEWS, REDDIT, WIKIPEDIA
  - WE INCLUDE 100 SENTENCE PAIRS CONSTRUCTED FROM EACH SOURCE AND 150 ARTIFICIALLY-CONSTRUCTED SENTENCE PAIRS FOR 550 TOTAL.
- ANNOTATION PROCESS - 42% ENTAILMENT, 35% NEUTRAL, AND 23% CONTRADICTION
- EVALUATION
  - MATTHEWS CORRELATION COEFFICIENT
  - ACCURACIES - 32.7% AND 36.4% RESPECTIVELY FOR SNLI AND MNLI
  - NLP RESEARCHERS ANNOTATE 50 SENTENCE PAIRS (100 ENTAILMENT EXAMPLES)
  - FLEISS'S 0.73.
  - THE AVERAGE R3 SCORE IS 0.80

BASELINES

- ARCHITECTURE
  - SENTENCE-TO-VECTOR ENCODERS
  - THE CLASSIFIER IS AN MLP WITH A 512D HIDDEN LAYER.
- PRE-TRAINING
  - ELMO - A PAIR OF TWO-LAYER NEURAL LANGUAGE MODELS (ONE FORWARD, ONE BACKWARD) TRAINED ON THE BILLION WORD BENCHMARK
  - COVE - A SEQUENCE TO SEQUENCE MODEL WITH A TWO-LAYER BILSTM ENCODER TRAINED FOR ENGLISH-TO-GERMAN TRANSLATION
- TRAINING
  - WE TRAIN OUR MODELS WITH THE BILSTM SENTENCE ENCODER AND POST-ATTENTION BILSTMS SHARED ACROSS TASKS, AND CLASSIFIERS TRAINED SEPARATELY FOR EACH TASK
- SENTENCE REPRESENTATION MODEL
  - GLOVE EMBEDDINGS (CBOW), SKIP-THOUGHT, INFERSENT, DISSENT AND GENSEN

BENCHMARK RESULTS

- MULTI-TASK TRAINING OVER SINGLE-TASK TRAINING YIELDS BETTER OVERALL SCORES, PARTICULARLY AMONG THE PARAMETER-RICH ATTENTION MODELS.
- ATTENTION GENERALLY HURTS PERFORMANCE IN SINGLE TASK TRAINING, BUT HELPS IN MULTI-TASK TRAINING
- ELMO EMBEDDINGS BETTER THAN GLOVE OR COVE EMBEDDINGS, PARTICULARLY FOR SINGLE-SENTENCE TASKS.
- USING COVE SLIGHTLY IMPROVES ON GLOVE FOR SINGLE TASK TRAINING BUT NOT FOR MULTI-TASK TRAINING.
- CONSISTENT GAINS BY MOVING FROM CBOW TO SKIP-THOUGHT TO INFERSENT AND GENSEN
- SOLVING GLUE IS BEYOND THE CAPABILITIES OF CURRENT MODELS AND METHODS, AND THAT TRAINING ON AUXILIARY TASKS SEEMS A NECESSARY AND PROMISING DIRECTION.

ANALYSIS

- COARSE CATEGORIES - OVERALL PERFORMANCE IS LOW FOR ALL MODELS:
  - THE HIGHEST TOTAL SCORE OF 28 STILL DENOTES POOR ABSOLUTE PERFORMANCE.
  - PERFORMANCE TENDS TO BE HIGHER ON PREDICATE-ARGUMENT STRUCTURE AND LOWER ON KNOWLEDGE
- FINE-GRAINED SUBCATEGORIES - MOST MODELS HANDLE UNIVERSAL QUANTIFICATION RELATIVELY WELL.
  - DOUBLE NEGATION IS ESPECIALLY DIFFICULT FOR THE GLUE-TRAINED MODELS THAT ONLY USE GLOVE EMBEDDINGS.
  - MODELS WERE SENSITIVE TO HYPERNYM/HYPONYM SUBSTITUTIONS AS SIGNALS OF ENTAILMENT, BUT PREDICTED IT IN THE WRONG DIRECTION

CONCLUSION

- IN AGGREGATE, MODELS TRAINED JOINTLY ON OUR TASKS SEE BETTER PERFORMANCE THAN THE COMBINED PERFORMANCE OF MODELS TRAINED FOR EACH TASK SEPARATELY.
- WE CONFIRM THE UTILITY OF ATTENTION MECHANISMS AND TRANSFER LEARNING METHODS SUCH AS ELMO IN NLU SYSTEMS, WHICH COMBINE TO OUTPERFORM THE BEST SENTENCE REPRESENTATION MODELS ON THE GLUE BENCHMARK