CSE 576: Topics in Natural Language Processing

Dataset - 3

Phase 2 Report

Task Statement: Understand the task and think about alternative ways to represent the same task. To create a variant task, you can write alternate definitions by changing writing style, vocabulary, length of definition etc. and change the instances generated by the dataset

Solution: We created 1-5 variants of the task definitions by paraphrasing the original definitions.

Created by: Anshita Singh Bais, Prutha Gaherwar

List of tasks:
task072_abductivenli_answer_generation.json
task073_commonsenseqa_answer_generation.json
task074_squad1.1_question_generation.json
task075_squad1.1_answer_generation.json
task076_splash_correcting_sql_mistake.json
task077_splash_explanation_to_sql.json
task078_splash_sql_to_explanation.json
task079_conala_concat_strings.json
task080_piqa_answer_generation.json
task081_piqa_wrong_answer_generation.json
task082_babi_t1_single_supporting_fact_question_generation.json
task083_babi_t1_single_supporting_fact_answer_generation.json
task084_babi_t1_single_supporting_fact_identify_relevant_fact.json
task085_unnatural_addsub_arithmetic.json
task086_translated_symbol_arithmetic.json
task087_new_operator_addsub_arithmetic.json
task088_identify_typo_verification.json
task089_swap_words_verification.json
task090_equation_learner_algebra.json
task092_check_prime_classification.json
task093_conala_normalize_lists.json
task094_conala_calculate_mean.json
task095_conala_max_absolute_value.json
task096_conala_list_index_subtraction.json
task097_conala_remove_duplicates.json

task098_conala_list_intersection.json
task102_commongen_sentence_generation.json
task103_facts2story_long_text_generation.json
task104_semeval_2019_task10_closed_vocabulary_mathematical_answer_generation.json
task105_story_cloze-rocstories_sentence_generation.json
task106_scruples_ethical_judgment.json
task107_splash_question_to_sql.json
task108_contextualabusedetection_classification.json
task109_smsspamcollection_spamsmsdetection.json
task110_logic2text_sentence_generation.json
task111_asset_sentence_simplification.json
task112_asset_simple_sentence_identification.json
task113_count_frequency_of_letter.json
task114_is_the_given_word_longest.json
task115_help_advice_classification.json
task116_com2sense_commonsense_reasoning.json
task117_spl_translation_en_de.json
task118_semeval_2019_task10_open_vocabulary_mathematical_answer_generation.json
task119_semeval_2019_task10_geometric_mathematical_answer_generation.json
task119_zest_text_modification.json
task120_zest_text_modification.json

For Tasks 72-95:

Number of variants created per task -
5
Tools used -
DiverseParaphraser
Casual2Former
Former2Casual
Quillbot.com
Styleformer
MySecondTransformation
Number of Instances Generated -
In the range 200 - 6500
New Instances generated for -
Task 72, 73, 74, 75, 80, 81
Tasks for which the instances were shuffled as all the instances of the dataset were used -
Tasks 76, 77, 78, 79, 82, 83, 84, 85, 86, 87, 88, 89, 90, 92, 93, 94, 95

Datasets Used -
Abductive NLI
CommonsenseQA
SQuad1.1

Splash
Conala
PIQA
BABI
Synthetic

Types of Definitions Generated-
Paraphrase, Long, Short, Formal, Casual, Summary

For Tasks 96-120:

Number of variants created per task - 1
Tools used -
sentence_reordering
style_paraphraser
concat_monolingual
styleformer
New instances were generated for the tasks -
102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 115, 116, 117, 118, 119, 119, 120
New instances couldn't be generated for the tasks due to lesser size of the training dataset -
96, 97, 98, 113, 114
Number of instances generated -
In the range 150 to 2500
Datasets used -
CoNala
commongen
Fact2Story
semeval 2019 task10
story cloze and ROCStories
scruples
splash
context abuse detection
SMS spam collection v.1
Logic2Text
Synthetic dataset
Help dataset from EMNLP 2020 paper
SPL
zest
asset
Com2Sense
Types of Definitions Generated-
Paraphrase, Long, Short, Formal, Casual, Summary