LogiQA 2.0—An Improved Dataset for Logical Reasoning in Natural Language Understanding

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang, *Member, IEEE*

***Abstract*—NLP research on logical reasoning regains momentum with the recent releases of a handful of datasets, notably LogiQA and Reclor. Logical reasoning is exploited in many probing tasks over large Pre-trained Language Models (PLMs) and downstream tasks like question-answering and dialogue systems. In this article, we release LogiQA 2.0. The dataset is an amendment and re-annotation of LogiQA in 2020, a large-scale logical reasoning reading comprehension dataset adapted from the Chinese Civil Service Examination. We increase the data size, refine the texts with manual translation by professionals, and improve the quality by removing items with distinctive cultural features like Chinese idioms. Furthermore, we conduct a fine-grained annotation on the dataset and turn it into a two-way natural language inference (NLI) task, resulting in 35 k premise-hypothesis pairs with gold labels, making it the first large-scale NLI dataset for complex logical reasoning. Compared to Question Answering, Natural Language Inference excels in generalizability and helps downstream tasks better. We establish a baseline for logical reasoning in NLI and incite further research.**

***Index Terms*—Reading comprehension, logical reasoning, natural language inference, textual inference.**

Premise: Met my first girlfriend that day.
Hypothesis: I didn't meet my first girlfriend until later.
Label: Contradiction

(a) An NLI example from the MNLI [11] dataset.

Passage: In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

Question 1: *What causes precipitation to fall?*
Answer: gravity
Question 2: *What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?*
Answer: graupel
Question 3: *Where do water droplets collide with ice crystals to form precipitation?*
Answer: within a cloud

(b) An MRC example from the SQuAD [12] dataset.

Fig. 1. Examples of traditional NLU benchmarks.

I. Introduction

THE capability of logical reasoning is a crucial part of natural language understanding (NLU) [1], [2], [3]. Investigation of *linguistic reasoning* dates back to the 1950 s, at the dawn of computer science and artificial intelligence [4], [5], [6], [7], [8]. However, with limited computing power and primitive NLU technologies, formal logical reasoning gradually dominated the research field in the 1970 s and became a key area of AI research over a long period [9], [10].

Recently, with the advance of deep learning technology, NLU has witnessed significant improvements [13], [14], with competitive results being reported over typical tasks, including natural language inference (NLI) [15], [16] and machine reading comprehension (MRC) [17], [18]. Fig. 1 illustrates the two NLU tasks. In Fig. 1(a), an NLI model takes the premise and hypothesis as input and predicts whether the premise entails the hypothesis. In Fig. 1(b), an MRC model takes a passage and question pair as input to predict the correct answer. There is a fundamental connection between machine reading comprehension and natural language inference [15], both tasks rely heavily on reasoning skills, and both are general because many NLP tasks can be cast into MRC [19], [20] or NLI [21]. For both NLI and MRC tasks, the current state-of-the-art approaches make use of a sizeable pre-trained language model such as BERT [13], and RoBERTa [22], fine-tuned using the benchmark-specific training data. Benefiting from large-scale pre-training, such models have achieved performances close to or surpass the human level on popular benchmarks [23], [24].

The recent advance in NLU leads to the natural question of whether it is time to revisit traditional *linguistic reasoning* tasks. Relevant to this question, some work has shown evidence that the current deep learning technologies have the potential to

Manuscript received 10 May 2022; revised 7 January 2023 and 25 April 2023; accepted 21 June 2023. Date of publication 6 July 2023; date of current version 9 August 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nancy F. Chen. *(Corresponding author: Yue Zhang.)*

Hanmeng Liu and Leyang Cui are with the Zhejiang University, Hangzhou 310007, China, and also with the School of Engineering, Westlake University, Hangzhou 310024, China (e-mail: liuhanmeng@westlake.edu.cn; cuileyang@westlake.edu.cn).

Jian Liu is with the Fudan University, Shanghai 200433, China, and also with the School of Engineering, Westlake University, Hangzhou 310024, China (e-mail: liujian@westlake.edu.cn).

Zhiyang Teng is with the School of Engineering, Westlake University, Hangzhou 310024, China (e-mail: tengzhiyang@westlake.edu.cn).

Nan Duan is with the Microsoft Research Asia, Beijing 100080, China (e-mail: nanduan@microsoft.com).

Ming Zhou is with the Langboat Technology, Beijing 100080, China (e-mail: zhouming@chuangxin.com).

Yue Zhang is with the School of Engineering, Westlake University, Hangzhou 310024, China, and also with the Institute of Advanced Technology, Westlake Institute of Advanced Study, Hangzhou 310024, China (e-mail: yue.zhang@wias.org.cn).

Digital Object Identifier 10.1109/TASLP.2023.3293046