Natural Language Inference in Context — Investigating Contextual Reasoning over Long Texts

Hanmeng Liu,[1] **Leyang Cui,**[1] **Jian Liu,**[2] **Yue Zhang**[3]
[1] Zhejiang University, [2] Fudan University, [3] Westlake University
liuhanmeng@zju.edu.cn, cuileyang@westlake.edu.cn, jianliu17@fudan.edu.cn, yue.zhang@wias.org.cn

Abstract

Natural language inference (NLI) is a fundamental NLP task, investigating the entailment relationship between two texts. Popular NLI datasets present the task at sentence-level. While adequate for testing semantic representations, they fall short for testing contextual reasoning over long texts, which is a natural part of the human inference process. We introduce ConTRoL, a new dataset for **ConT**extual **R**easoning **o**ver **L**ong Texts. Consisting of 8,325 expert-designed "context-hypothesis" pairs with gold labels, ConTRoL is a passage-level NLI dataset with a focus on complex contextual reasoning types such as logical reasoning. It is derived from competitive selection and recruitment test (verbal reasoning test) for police recruitment, with expert level quality. Compared with previous NLI benchmarks, the materials in ConTRoL are much more challenging, involving a range of reasoning types. Empirical results show that state-of-the-art language models perform by far worse than educated humans. Our dataset can also serve as a testing-set for downstream tasks like checking the factual correctness of summaries.

Introduction

Natural languages are powerful tools for reasoning. In NLP, natural language inference (NLI) has attracted surging research interests (Bowman et al. 2015; Williams, Nangia, and Bowman 2018; Bhagavatula et al. 2020). The task is to determine whether a hypothesis h can reasonably be inferred from a premise p. Thanks to the generalizability of the NLI framework (i.e., nearly all questions about meaningfulness in language can be reduced to questions of entailment and contradiction in context), NLI can serve as a proxy to general tasks such as natural language understanding (NLU). As a result, the NLI task is constantly employed as a testing ground for learning sentence representation as well as evaluating language models, with the expectation of benefiting downstream applications.

Large-scale NLI datasets have been collected via crowdsourcing. Existing benchmarks (Bowman et al. 2015; Williams, Nangia, and Bowman 2018; Dagan, Glickman, and Magnini 2005; Khot, Sabharwal, and Clark 2018a) handle the task at the sentence-level, generating labelled sentence pairs by probing into the essence of lexical and compositional semantics. These benchmarks explore rich features of sentence meaning, testing various aspects of semantic representation. With the advance of contextualized embeddings such as BERT (Devlin et al. 2019), pre-trained language models achieve competitive results. The state-of-the-art models can even reach human-level performance.

P: Ten new television shows appeared during the month of September. Five of the shows were sitcoms, three were hour-long dramas, and two were news-magazine shows. By January, only seven of these new shows were still on the air. Five of the shows that remained were sitcoms.

H1: *At least one of the shows that were cancelled was an hour-long drama.*
Entailment ✓ Contradiction Neutral

H2: *There is no hour-long drama remained on the air.*
Entailment **Contradiction** ✓ Neutral

H3: *Television viewers prefer sitcoms over hour-long dramas.*
Entailment Contradiction **Neutral** ✓

Figure 1: An example of the ConTRoL dataset. (✓ indicates the correct answer.)

Contextual reasoning is essential to the process of human cognition, where inference is made based on contextual information and a collection of facts (Giunchiglia 1992). Inferring hidden facts from context is an indispensable element of human language understanding. Contextual reasoning is typically performed on the passage level, where multiple steps may be necessary for inferring facts from given evidences. It has been investigated by NLP tasks such as machine reading (Lai et al. 2017; Sun et al. 2019a), retrieval-based dialogue (Wu et al. 2017). However, dominant NLI benchmarks (Bowman et al. 2015; Williams, Nangia, and Bowman 2018) investigate the relationship of two sentences, with relatively less attention being payed to the exploration of grounded logical inference (Bhagavatula et al. 2020; Clark, Tafjord, and Richardson 2020).

We investigate contextual reasoning for NLI by making a dataset that consists of 8,325 instances. One example is shown in Figure 1. In this example, the premise consists of several facts concerning a set of shows, which can serve as a context for evidence integration and reasoning. The truthfulness of the hypotheses are determined by reasoning over multiple sentences. Various types of contextual reason-