MSCINLI: A Diverse Benchmark for Scientific Natural Language Inference

The paper "MSCINLI: A Diverse Benchmark for Scientific Natural Language Inference" by Mobashir Sadat and Cornelia Caragea, published in *arXiv e-prints* in April 2024, introduced MSCINLI, a dataset containing 132, 320 sentence pairs extracted from five new scientific domains.

Despite introducing a challenging task and enabling the exploration of NLI with scientific text, SCINLI lacks the diversity to serve as a general purpose scientific NLI benchmark because it is limited to a single domain (ACL). Moreover, due to the unavailability of multiple domains, SCINLI is not suitable for studying domain adaptation and transfer learning on scientific NLI. To this end, in this paper, they propose MSCINLI, a scientific NLI dataset containing 132, 320 sentence pairs extracted from papers published in five different domains: "Hardware", "Networks", "Software & its Engineering", "Security & Privacy", and "NeurIPS."

They evaluate the difficulty of MSCINLI by experimenting with a BiLSTM based model. They then establish strong baselines on MSCINLI by a) fine-tuning four transformer based Pre-trained Language Models (PLMs): BERT, SCIBERT, ROBERTA and XLNET; and b) prompting two Large Language Models (LLMs) in both zero-shot and few-shot settings: LLAMA-2, and MISTRAL. Furthermore, they provide a comprehensive investigation into the robustness of scientific NLI models by evaluating their performance under domain-shift at test time. Finally, they explore both SCINLI and MSCINLI in an intermediate task transfer learning setting to evaluate their usefulness in improving the performance of other downstream tasks.

Their key findings include: a) MSCINLI is more challenging than SCINLI; b) the best performing PLM baseline, which is based on ROBERTA, shows a Macro F1 of 77.21% on MSCINLI indicating the challenging nature of the task and a substantial headroom for improvement; c) the best performing LLM baseline with LLAMA-2 shows a Macro F1 of only 51.77% indicating that their dataset can be used to benchmark the NLU and complex reasoning capabilities of powerful LLMs; d) domain-shift at test time reduces the performance; and e) diversity in the scientific NLI datasets helps to improve the performance of downstream tasks.