SCINLI: A Corpus for Natural Language Inference on Scientific Text

The paper "SCINLI: A Corpus for Natural Language Inference on Scientific Text" by Mobashir Sadat and Cornelia Caragea, published in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* in May 2022, introduced SciNLI, a large dataset for NLI that captures the formality in scientific text and contains 107, 412 sentence pairs extracted from scholarly papers on NLP and computational linguistics.

Although a lot of natural language inference datasets are available, none of the existing NLI datasets is related to scientific text that is found in research articles. The vocabulary as well as the structure and formality used in sentences in scientific articles are very different from the sentences used in the everyday language. Thus, the authors decide to introduce SciNLI.

To capture the inference relations which are prevalent in scientific text but are unavailable in the existing NLI datasets, they introduce two new classes — CONTRASTING and REASONING. Because of the differences in the frequency of occurrence of the linking phrases related to different classes, their initial dataset was unbalanced in all three splits, they balance their dataset by downsampling the top three most frequent classes to the size of the least frequent class in each split. They construct their training set from scientific papers on NLP and computational linguistics available in the ACL Anthology, published between 2000 and 2019. They create SCINLI by harnessing cues in their data in the form of linking phrases between contiguous sentences, which are indicative of their semantic relations and provide a way to build a labeled dataset using distant supervision. During training, they directly utilize these (potentially noisy) sentence pairs, but to ensure a realistic evaluation of the NLI models over scientific text, they manually annotate 6,000 sentence pairs. These clean pairs are used in two splits, 2,000 pairs for development and hyperparameter tuning and 4,000 pairs for testing. Similar to the related datasets, they parse the sentences in SciNLI by using the Stanford PCFG Parser (3.5.2). We can see that $\approx 97\%$ of both first and second sentences have parses with an 'S' root which

is higher than the sentences in SNLI and very competitive with the other datasets, reflecting that most of the sentences are syntactically complete.

They evaluate SCINLI by experimenting with traditional machine learning models using lexical and syntactic features, neural network models— BiLSTM, CBOW, CNN, and pre-trained language models—BERT, SciBERT, RoBERTa, and XLNet. Their findings suggest that: (1) SCINLI is harder to classify than other datasets for NLI; (2) Lexical features are not enough for a model to achieve satisfactory performance on SCINLI and deep semantic understanding is necessary; (3) SCINLI is well suited for evaluating scientific NLI models; and (4) Their best performing model based on XLNet shows 78.18% Macro F1 and 78.23% accuracy illustrating that SCINLI is a challenging new benchmark.