FRI Natural language processing course 2023/24
Project report

University *of Ljubljana*
Faculty *of Computer and*
Information Science

Towards Generalizable NLP Models: A Prompt-Guided Inference Dataset with Integrated Reasoning Rules

Blaž Lipar and Meta Kokalj

Abstract

Keywords
natural language inference (NLI), reasoning, prompt engineering, inference scenario creation, paragraph-level inference

Advisors: Aleš Žagar

Introduction

Natural Language Inference (NLI) tasks serve as a valuable benchmark for assessing the capabilities of new models within Natural Language Understanding (NLU) research. This is particularly significant as NLU plays a crucial role in the success of many actively studied NLP problems, including question answering, translation, and dialogue [1].

Interestingly, the field of Natural Language Processing (NLP) often overlooks the role of paragraphs despite their importance in discourse and text generation. The majority of NLI datasets and research focused only on sentence-level inference, where the premises and hypotheses are expressed in single, typically brief sentences. They are therefore inappropriate for use in other open-ended NLP situations. This is particularly crucial in tasks aimed at verifying the factual accuracy of summaries of documents [2]. Paragraphs, often described as ”units of thought,” go beyond the grammatical structure of individual sentences (syntax) [3]. Analyzing how sentences within a paragraph connect through cohesion (referential links) and coherence (logical flow of ideas) unlocks a deeper understanding of meaning [4, 5, 6]. This analysis of relationships between sentences is crucial for tasks like discourse generation, where sentences need to be arranged in a way that creates a clear and cohesive flow of information. Incorporating paragraph-level semantics into NLP models could significantly improve their ability to handle real-world language tasks.

In addition, many crowdsourced NLI datasets, while convenient, may not be well-suited for tackling real-world NLP problems due to their creation in isolation from specific downstream tasks and due to inherent annotator biases [7, 8]. Specifically, as shown in the case of Multi-Genre Natural Language Inference (MultiNLI), crowd workers sometimes relied on certain strategies to form hypotheses of a specific label like adding negators for contradiction labels, and introducing bias into the dataset. This approach creates an unrealistically easy task, failing to capture the true complexity of real-world human reasoning [8]. Many analyses of the SNLI dataset conducted in recent years have revealed that models trained on it often lack significant compositionality in their representations. Instead, these models primarily rely on straightforward heuristics, such as word overlap, when making inferences (this phenomenon is commonly recognized as the ”lexical overlap heuristic,” as articulated by McCoy et al.(2019) [9]. For instance, a high overlap in words between premises and hypotheses typically predicts entailment, while contradictory sentence pairs usually exhibit minimal or no word overlap and/or the presence of negation words. This underscores the importance of developing models that can effectively incorporate external knowledge for more nuanced natural language inference tasks [10, 11].

What is emerging NLI's most pressing problem is their propensity for overfitting. While these models demonstrate impressive performance on the datasets they are trained upon, potentially exceeding human benchmarks, their ability to generalize to unseen data remains severely limited. This phenomenon, well-established in machine learning, is referred to as the generalization problem [12]. An underlying cause might be the focus on deductive reasoning in training data creation. Careful examination of crowd worker instructions for popular NLI datasets reveals an emphasis on logically guaranteed inferences, neglecting inductive reasoning, which involves concluding unseen data.

To address these challenges and generate more realistic, nuanced data, we propose a novel paragraph scope scenario-inference pair elicitation method where each scenario is accompanied by multiple potential inferences. By incorporating various reasoning modes within the prompt creation stage, we aim to train an NLI model that can handle a broader spectrum of reasoning patterns, ultimately leading to a more robust and generalizable understanding of natural language inference.

Recent work

Pioneering datasets like The Stanford Natural Language Inference (SNLI) [13] and the Multi-Genre Natural Language Inference (MNLI) [1] established a benchmark with their large scale, balanced data, and clear labels. However, as the field matures, researchers are constantly innovating, recognizing limitations in existing datasets such as reliance on human-generated sentences and potential biases.

Cornerstone Datasets

The Stanford Natural Language Inference (SNLI) dataset, a cornerstone for NLI research, offers a large-scale, balanced dataset with clear labels, facilitating robust models for diverse logical relationships [13]. However, limitations exist. The simplistic sentence generation process from simple, static scenes limits reasoning style diversity, and the reliance on affirmative image captions restricts the ability to assess negation, a crucial aspect of natural language [14]. Building upon SNLI, the Multi-Genre Natural Language Inference (MNLI) dataset extends the scope by incorporating sentence pairs from various genres and domains (e.g., fiction, telephone conversations) [1]. This inclusion of diverse genres promotes model robustness by exposing them to a wider range of language styles and reasoning patterns. Additionally, MNLI facilitates transfer learning tasks, where models trained on this dataset can be adapted to perform well on different NLP applications. While MNLI offers significant advancements, it inherits limitations from SNLI, such as reliance on human-generated sentences, and suffers from genre imbalance within the dataset. Gururangan et al. [15] questioned whether SNLI-trained models are achieving genuine language comprehension or simply capitalizing on coincidental patterns (artifacts) within the dataset.

Domain-Specific and Challenging Datasets

There are many existing NLI training datasets, each with its strengths and weaknesses. SciTail [7]focuses on scientific language, training models for reasoning specific to scientific contexts and encouraging understanding of contextual relevance but is limited in size. FEVER (Fact Extraction and VERification) tackles real-world fact-checking challenge with claims and evidence from Wikipedia. While valuable for these tasks, FEVER can be susceptible to biases and noise from imperfect evidence retrieval [**?**]. BioNLI is a dataset in biomedical natural language inference [16]. It's particularly interesting that this research addresses the automatic creation of meaningful negative examples by using rule-based and neural-based counterfactuals, potentially removing the need for expert involvement.

The development of NLI datasets is a continuous process. The paper of Liu et al. [17] introduced a new dataset called ConTRoL for contextual reasoning in natural language inference. It is a passage-level NLI dataset that focuses on complex reasoning types and is more challenging than previous benchmarks. Rather than evaluating a hypothesis versus a short passage, Koreeda and Manning [18] introduced evaluation of the hypothesis against a whole document. The system is provided with a contract and a set of hypotheses, which could be statements about the contract's obligations, rights, or other key points. This approach, applied to contracts, requires models to understand the broader context of the document. Worth noting is the study by Nie et al. (2020) which presented a large-scale NLI benchmark dataset collected through an iterative human-in-the-loop approach, resulting in a dataset that surpasses existing benchmarks in difficulty. Beyond the dataset itself, the authors proposed a novel data collection method that functions as a "moving target" for NLU. Unlike static benchmarks that models can quickly overfit on, this approach acts as a never-ending learning scenario that allows for continuous improvement and adaptation [19].

HANS (Heuristic Analysis for NLI Systems) [20] and ANLI (Adversarial NLI) [19] both aim to improve the robustness of NLI models. HANS achieves this by exposing model biases through pre-defined templates, while ANLI uses specially crafted adversarial examples.

Parlay creation

This project seeks to construct a comprehensive inference dataset for training a NLP model. The proposed dataset (Parlay) will be designed to promote model generalizability by encompassing five distinct categories, ranging from historical events to fictional narratives as inference pairs. To ensure the dataset's further diversity, each category will include inference pairs drawn from a wide range of subfields within its domain. Clear and concise prompts will be employed to introduce scenarios within each category. These scenarios will provide a brief description of a situation or event, offering essential context for the model to perform various inference tasks effectively.

The selection of scenario categories – historical events, scientific explanations, everyday life situations, news reports, and fictional stories – is strategically chosen to provide a comprehensive training ground for various reasoning modes. Each category offers unique advantages. Historical Events: Examining historical events allows the model to analyze complex cause-and-effect relationships, temporal reasoning (e.g., order of events), and reasoning about counterfactuals ("what if" scenarios) [21, 22, 23, 24, 25, 26, 27]. Scientific Explanations: Scientific explanations necessitate the understanding of deductive logical reasoning rules like modus ponens and modus tollens [28]. These scenarios will train the model to identify key evidence and draw valid conclusions based on established scientific principles. Everyday situations: Everyday scenarios provide a fertile environment for practical reasoning. The model encounters incomplete information, necessitating inductive reasoning based on patterns, experiences, and common sense [29]. Additionally, everyday scenarios require analogical reasoning, where the model identifies similarities to use knowledge from one situation to understand another [30]. This develops transferable reasoning skills, allowing the model to adapt to novel situations. On top of that, informal language plays a significant role in everyday situations. As everyday life moves more and more into a virtual environment, understanding the diverse system of informal language is of great importance. Examples of everyday conversations in informal language will be included to cover the wide range of inputs and forms, in which informal language appears. News Reports: News reports naturally require the model to exercise critical reasoning skills [31]. This involves evaluating source credibility, identifying potential biases, and distinguishing factual information from opinion. Navigating these complexities equips the model to make sound judgments based on the presented information. Fiction: Fictional stories provide a platform for exploring narrative reasoning, including understanding characters' motivations, emotions (theory of mind) [32, 33]. Furthermore, because stories often involve social interactions and dynamics the model can learn to reason about how characters behave in social contexts and predict their actions based on social norms (social reasoning) [34, 35].

We opted for an alternative approach utilizing prompt-driven scenario generation. While inspired by natural texts and ideas as prompt inputs, the scenarios themselves will be newly created based on those prompts. This approach addresses key challenges inherent in natural text-based datasets: citation management and authorship attribution difficulties. Academic and ethical integrity necessitate proper citation of any scientific articles or other source material used in a dataset. Including complete citations within the dataset itself might not be feasible due to space constraints or the chosen format (e.g., plain text pairs). Manually creating and managing citations for a large dataset can also be incredibly time-consuming. Prompt-driven scenarios eliminate this burden. Natural texts often lack clear authorship, making it difficult to assess potential biases. The collaborative nature of news production, frequent use of anonymous sources, and information aggregation approach in books make it difficult to trace the origin of specific ideas. Additionally, journalists' interpretations and re-presentations of existing information can obscure the original author of underlying concepts. Unclear authorship makes it difficult to assess the potential biases associated with the information, and when creating inference pairs from different fields, you might inadvertently introduce biases based on the original authors' viewpoints. This can ultimately impact the credibility of datasets constructed from natural texts. Our prompts allow for greater control over the information presented, mitigating bias.

To be concrete, we intend to construct a comprehensive NLI (Natural Language Inference) dataset, using the capabilities of large language models (LLMs) for generating paragraph pairs. Each instance within our dataset will encompass three outputs, each comprising two paragraphs. These paragraphs will be categorized into three distinct relations: entailment, neutral, or contradiction. We have chosen Slovenian as the language for crafting these paragraphs, ensuring linguistic diversity and relevance. To facilitate the LLMs' understanding and alignment with our dataset's structure, we will initially introduce them to the intricacies of the SI-NLI dataset. This introduction will serve to acquaint the LLMs with the precise parameters and expectations of our inquiry, enabling them to produce outputs that adhere closely to our intended format and objectives. In order to understand the generation process better, we will augment it by providing explicit rationales behind the LLMs' decisions in creating specific paragraph pairs. This involves the explanation of the LLMs underlying logic and reasoning that inform the categorization of each pair into entailment, neutral, or contradiction. By giving us feedback on the decision-making process, we aim to enhance the interpretability and general quality of the generated outputs. Furthermore, to elevate the diversity and quality of the generated outputs, we will implement measures aimed at the creation of redundant or highly similar tokens between the premise and hypothesis. This entails ensuring that each pair of paragraphs exhibits sufficient variance in language and content, thereby enriching the dataset with a wider spectrum of linguistic patterns and semantic nuances.

By utilizing prompt-driven scenarios, we will ensure the creation of a high-quality and well-controlled foundation for generating diverse inference pairs. We will achieve this through the following methods:

Reasoning Rule-Guided Prompt Engineering. We will develop prompts that incorporate established reasoning rules, such as modus ponens, modus tollens, causal reasoning, textual entailment, and predictive inference (Liu, Teng, Ning, Liu, Zhou, Zhang, 2023). Our focus will be on crafting effective prompts that integrate factual hints and guide the model towards specific inferences (entailment, contradiction, neutral). Factual hints refer to pieces of information embedded within the prompts that provide the model with essential knowledge about the scenario. These hints serve as anchors for the model's reasoning process, allowing it to make inferences based on established facts. While factual hints provide a strong foundation, real-world inferences often rely on broader reasoning patterns beyond established rules. To address this, we will incorporate heuristics within our prompts. Heuristics can capture these general reasoning patterns, allowing our prompts to be more adaptable to diverse scenarios and inferences encountered in real-world situations. We will achieve this by exploring templates or conditional statements within the prompts based on the reasoning rule we want to target. Analyzing existing NLP datasets like SNLI and MNLI will be crucial to understand how prompts can be structured for different inferences."

Chain of Thought Reasoning. We will investigate the application of chain of thought reasoning, where the model explicitly reveals its reasoning steps as it arrives at an inference. This will allow us to gain deeper insights into the model's decision-making process, especially in relation to the applied reasoning rules, and further refine our prompt design for optimal performance [36].

Tree of Thought (ToT) Technique. Additionally, we will explore the recently proposed Tree of Thought (ToT) technique [37]. This technique encourages the model to explore different reasoning paths and evaluate them before arriving at a conclusion. By studying how ToT influences the model's reasoning process, we can potentially improve the quality and diversity of generated inference pairs.

We will detail the specifics of this approach and its advantages in the following section.

Results

Discussion

Acknowledgments

References

[1] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2018.

[2] Wenpeng Yin, Dragomir Radev, and Caiming Xiong. Docnli: A large-scale dataset for document-level natural language inference. *arXiv preprint arXiv:2106.09449*, 2021.

[3] Wlodek Zadrozny and Karen Jensen. Semantics of paragraphs. *Computational Linguistics*, 17(2):171–210, 1991.

[4] Hesham Suleiman Alyousef. Text cohesion in english scientific texts written by saudi undergraduate dentistry students: A multimodal discourse analysis of textual and logical relations in oral biology texts. *Sage Open*, 11(3), 2021.

[5] Wolfram Bublitz. Discursive pragmatics. In Jan Zienkowski, Jan-Ola Östman, and Jef Verschueren, editors, *Handbook of Pragmatics Highlights 8*, pages 37–49. 2011.

[6] Advaith Siddharthan. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109, 2006.

[7] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *Proceedings of AAAI*, pages 5189–5197, 2018.

[8] Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Lawrence S. Moss. Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*, 2020.

[9] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

[10] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *arXiv preprint arXiv:1812.01193*, 2018.

[11] Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*, 2018.

[12] Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks. *Journal of Logic, Language and Information*, 33:21–48, 2024.

[13] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[14] Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. An analysis of natural language inference benchmarks through the lens of negation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online, November 2020. Association for Computational Linguistics.

[15] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[16] Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. Bionli: Generating a biomedical nli dataset using lexico-semantic constraints for adversarial examples. *arXiv preprint arXiv:2210.14814*, 2022.

[17] Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. Natural language inference in context – investigating contextual reasoning over long texts. *arXiv preprint arXiv:2011.04864*, 2020.

[18] Yuta Koreeda and Christopher Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In Marie-Francine Moens, Xuanjing Huang,

Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[19] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics.

[20] Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. Generating data to mitigate spurious correlations in natural language inference datasets. *arXiv preprint arXiv:2203.12942*, 2022.

[21] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[22] Marta Tatu and Munirathnam Srikanth. Experiments with reasoning for temporal relations between events. In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference*, Manchester, UK, August 2008.

[23] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems. *arXiv preprint arXiv:1209.2355*, 2013.

[24] Lucius E. J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. A new paradigm for counterfactual reasoning in fairness and recourse. *arXiv preprint arXiv:2401.13935*, 2024.

[25] Zhihong Deng, Jing Jiang, Guodong Long, and Chengqi Zhang. Causal reinforcement learning: A survey. *arXiv preprint arXiv:2307.01452*, 2023.

[26] Roman Abel and Martin Hänze. Generating causal relations in scientific texts: The long-term advantages of successful generation. *Frontiers in Psychology*, 10:199, 2019.

[27] Jie Yang, Soyeon Caren Han, and Josiah Poon. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64:1161–1186, 2022.

[28] Ronald J. Brachman and Hector J. Levesque. *Knowledge Representation and Reasoning*. Elsevier, 2004.

[29] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[30] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[31] Danjie Chen, Yan Zheng, Liqun Ma, and Fen Qin. An ontology-based knowledge representation method for typhoon events from chinese news reports. *ISPRS International Journal of Geo-Information*, 12(9):383, 2023.

[32] Renate LEP Reniers, Rhiannon Corcoran, Richard Drake, Nick M Shryane, and Birgit A Völlm. The qcae: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, 93(1):84–95, 2011.

[33] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY, 2011.

[34] Debjit Paul and Anette Frank. Social commonsense reasoning with multi-head knowledge attention. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2969–2980, Online, November 2020. Association for Computational Linguistics.

[35] Peng Wu, Xiaotong Li, Chen Ling, Shengchun Ding, and Si Shen. Sentiment classification using attention mechanism and bidirectional long short-term memory network. In *Proceedings of the xxx International Conference on Artificial Intelligence (AAAI)*, 2018.

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2023.

[37] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.