INFERENCE AND REASONING FOR SEMI-STRUCTURED TABLES

by

Vivek Gupta

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

School of Computing

The University of Utah

May 2023

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Vivek Gupta**
has been approved by the following supervisory committee members:

Vivek Srikumar ,	Chair(s)	**31 March 2023** Date Approved
Jeffrey Phillips ,	Member	**31 March 2023** Date Approved
William Wang ,	Member	**1 April 2023** Date Approved
Mohit Bansal ,	Member	**1 April 2023** Date Approved
Ellen M. Riloff ,	Member	**3 April 2023** Date Approved

by **Mary W. Hall** , Chair/Dean of
the Department/College/School of **Computing**
and by **David B. Kieda** , Dean of The Graduate School.

ABSTRACT

Semi-structured tabular data, such as ones in e-commerce product descriptions, annual financial reports, sports score statistics, scientific articles, etc., are ubiquitous in real-world applications. This dissertation investigates how machines understand and reason about such data. Understanding the meaning of text fragments and their implicit connections is essential for processing such data.

The author introduces the INFOTABS dataset, which presents a challenge for traditional modeling techniques due to its semi-structured, multi-domain, and heterogeneous nature. To overcome these challenges, effective ways of incorporating knowledge into reasoning models are explored. This approach involves using simple pre-processing strategies and leveraging structured data knowledge graphs. Additionally, the author proposes a cost-effective pipeline for translating tables to address the challenge of multilingual tabular inference, which enables the extension of INFOTABS to a multilingual version called XINFOTABS.

Through systematic probing, it was observed that existing models do not reason with tabular facts despite accurate predictions. Therefore, a trustworthy tabular inference approach involving two-stage evidence extraction and inference prediction was proposed. Additionally, semi-automatic data augmentation techniques were investigated, and the AUTO-TNLI dataset was introduced to improve reasoning on the INFOTABS dataset. To further enhance model robustness, a prompt-based learning approach was introduced that extracts knowledge from semi-structured tables, thus improving performance and robustness on adversarial tests.

The work opens up several new directions for future work involving reasoning on dynamic, multilingual, and multi-modal semi-structured tabular information.

To my parents and my sisters

CONTENTS

ACKNOWLEDGEMENTS

I am immensely grateful to all the individuals and organizations who have helped me throughout my Ph.D. journey. Your support, guidance, and encouragement have been invaluable and have contributed significantly to my academic and personal growth. First and foremost, I would like to express my deepest appreciation to my family for their unwavering love and support. I am grateful for my parents and sisters, who have always been my pillars of strength and motivation. Their constant encouragement and belief in me have been instrumental in my academic success.

I am indebted to my advisor, Prof. Vivek Srikumar, for his guidance, patience, and mentorship. I am deeply grateful for his unwavering support and mentorship throughout my Ph.D. journey. His extensive knowledge of natural language processing and machine learning has been instrumental in shaping my research and enabling me to grow as a researcher. Prof. Srikumar's constant availability, patient listening, insightful feedback, and encouragement to explore new ideas while keeping me on track with the project's goals have been invaluable. He has also played a critical role in shaping my research interests and career aspirations by introducing me to potential collaborators and recommending me for various opportunities. I am also thankful to my committee members Prof. Jeff Phillips, Prof. Ellen Riloff, Prof. William Wang, and Prof. Mohit Bansal for their valuable feedback, insightful discussions, and support throughout my Ph.D. journey. Their expertise in various aspects of computer science and natural language processing has enriched my research and made it more robust.

I would like to express my sincere gratitude to Bloomberg LP. for the Bloomberg Data Science Ph.D. Fellowships (2021-2023) and especially to my Bloomberg Ph.D. fellowship mentor, Dr. Shuo Zhang, for his constant support, guidance, and enlightening discussions. Dr. Zhang has been an invaluable mentor, providing timely feedback and guidance on research and paper publication. His dedication to mentoring has inspired me to become a better researcher. This fellowship has been transformative, and I look forward to applying

what I've learned to future research projects. Thanks to Dr. Anju Kambadur (Head AI, Bloomberg LP.) for creating such industry academia collaboration opportunities.

Special thanks to Prof. Suresh Venkatasubramanian and Prof. Ellen Riloff for teaching me so much about research problem selection, the importance of rigor in research, and creating valuable research impact. I am grateful to Dr. Maneesh Singh, Manish Srivastava, and Julian Martin Eisenschlos from Verisk Inc, IIIT Hyderabad, and Google Research respectively. for his valuable support and guidance, especially on rigor on writing, and teaching me approach to brainstorm on complex and challenging problem and ideas. I would also like to thank Prof. Erin Parker and Prof. Daniel Kopta for giving me the opportunity to learn about teaching ans also practically teaching my one of my favorite topics (Graph Data Structure and Algorithms) at the University of Utah. Thanks to Utah Data Science Center Director Prof. Jeff Phillips, Associate Director for Research Prof. Aditya Bhaskara (2021), Prof. Blair Sullivan (2022), Prof. Suresh Venkatasubramanian (2021), and Associate Director for Student Engagement Prof. Vivek Srikumar (2020), Prof. Anna Little (2021), and Associate Director for Outreach Bei Wang Phillips for providing me with the opportunity to participate in the organization of Data Science Seminar, Utah Data Science Club Activities, and Utah Data Science Days. Special thanks to all members especially, Ana, Fateme, Atreya, Ashim, Yichu, Tao, Giorgi, Tarun, Ana, Maitrey, Tianyu, Mattia, Yuan of the Utah NLP Group for their valuable feedback.

During my Ph.D. journey, my friends played a crucial role in keeping me motivated, encouraged, and keeping me sane. I am incredibly grateful to have had such an amazing support system in my life. They were always there to lend a listening ear, offer advice, and cheer me up when I needed it the most. I especially want to thank my flatmates who have been with me throughout my entire Ph.D. journey. Ankit, Piyush, Mahesh, Arnab - these guys were more than just roommates, they were my family away from home. I am also grateful to my other special friends (many from the Badminton Group) who have been a source of inspiration and motivation for me. Pegah, Sumana, Tushar, Maitrey, Ashim, Amanpreet, Saurabh, Tripti, Pratishtha, Himani, Sreeja, Srabani, Ravi, Manila, Mahima, Nikita, Mahesh, Emin, Sevda, Sarabjeet, Ananth, and others - I cannot thank them enough for being there for me whenever I needed them. They shared my joys and my struggles, and we supported each other through thick and thin. From celebrating a successful paper

submission to pulling all-nighters during exam season, we went through it all together. They brought joy and laughter into my life and provided a much-needed break from the rigors of academic life. Also special thanks to my Spiritual friends (Anjan, Harsha, Himani, Rahul, Vidhusi and others.) and other Guru's (Prof. Ramesh Goel, Prof. Prashant Saraswat, Mrs. Anushree Goel) who lead me stay grounded and connected to God.

My Ph.D. journey would have been incomplete without the invaluable contributions of my student collaborators. Their willingness to work with me and their exceptional research capabilities have enabled me to take on more challenging research projects and achieve groundbreaking results. I am grateful to have had the opportunity to work with such a talented and motivated group of individuals who have inspired me with their ideas and dedication. In addition, I would like to express my gratitude to the wider NLP community for their valuable insights, support, and encouragement throughout my Ph.D. journey. The various journals and conferences, such as TACL, ACL, NAACL, EMNLP, AACL, COLING, etc., have been a constant source of knowledge, inspiration, and exposure to the latest advancements in NLP research. I am fortunate to have had the opportunity to attend these events and present my work to some of the most accomplished researchers in the field. To all others who have supported me during my Ph.D. journey and whom I have missed mentioning, thank you. Your contributions have been invaluable and have made this journey worthwhile.

I am grateful for the support provided by several funding agencies and organizations during the course of my research. Specifically, I acknowledge grants received from NSF, Google, University of Utah, and Verisk, which have enabled me to pursue my work in semi-structured tables. Additionally, I would like to express my gratitude for Ph.D. Fellowship from Bloomberg, which has contributed significantly to this research. Finally, I am grateful for the compute resources grants received from Microsoft and Google Collaboration, which have been instrumental in facilitating the experiments and computations required for this research. The lessons I have learned during my Ph.D. journey will undoubtedly shape my future research work. I am excited to apply the skills and knowledge I have gained to new and challenging research projects. With the support of my advisors, colleagues, and friends, I am confident that I can continue to make meaningful contributions to the NLP community and advance the state of the art in this exciting field.

CHAPTER 1

INTRODUCTION

Semi-structured tables are a ubiquitous feature in various domains, including ecommerce product listings, finance annual reports, sports score tables, scientific articles, etc. Despite their varied contexts, these tables share some common characteristics. One notable attribute is their succinct nature; they can hold a large amount of information in a compact form. Thus, making them an ideal tool for comparative analysis and finding information. Additionally, tables (such as the one in Figure 1.1) require complex reasoning and inference to understand the implicit connections across table cells.

Although neural network models have gained success on unstructured text (sentences and paragraphs), their reasoning capacity on semi-structured text is poorly understood. Consequently, people (even NLP experts) have limited perception on how models reason. Reasoning over semi-structured tabular text involves comprehension of the meaning of text fragments and implicit relationships between tabular entries. Thus, semi-structured data can serve as a crucial testing ground for increasing the understanding of how Natural Language Processing (NLP) models reason about information. Thus, studying semi-structured data is essential for understanding model reasoning ability on textual information. Therefore, this dissertation focuses on entity-centric *semi-structured tabular data*.

1.1 Reasoning about Semi-Structured Data

We often encounter textual information that is neither unstructured (i.e., raw text) nor strictly structured (e.g., databases). Such data, where a structured scaffolding is populated with free-form text, can range from the highly verbose (e.g., web pages) to the highly terse (e.g. fact sheets, information tables, technical specifications, material safety sheets). Unlike databases, such semi-structured data can be heterogeneous in nature, and not characterized by pre-defined schemas. Moreover, we may not always have accompanying explanatory text that provides context. Yet, we routinely make inferences about such

heterogeneous, incomplete information and fill in gaps in the available information using our expectations about relationships between the elements in the data.

Understanding semi-structured information requires a broad spectrum of reasoning capabilities. We need to understand information in an ad hoc layout constructed with elements (cells in a table) that are text snippets, form fields or are themselves sub-structured (e.g., with a list of elements). Querying such data can require various kinds of inferences. At the level of individual cells, these include simple lookup (e.g., knowing that *Breakfast in America is recorded in a studio*), to lexical inferences (e.g., understanding that *Length* means the total recording songs time for the album), to understanding types of text in the cells (e.g., knowing that the number 1979 is a year and 46:06 is in minutes). Moreover, we may also need to aggregate information across multiple rows (e.g., knowing that *comparison of release and recording date month for the album*), or perform complex reasoning that combines temporal information with world knowledge. A true test of reasoning should evaluate the ability to handle such semi-structured information.

1.2 Tabular Natural Language Inference

Natural Language Inference (NLI) is the task of determining if a hypothesis sentence can be inferred as true (ENTAIL), false (CONTRADICT), or undetermined (NEUTRAL) given a premise sentence [45]. Contextual sentence embeddings such as BERT [51], RoBERTa [158], and DeBERTa [92] applied to large NLI datasets such as SNLI [17] and MultiNLI [280], have led to near-human performance of NLI systems, on benchmarks such as Glue [270] and SuperGlue [269]. In this dissertation, we study this question by proposing an extension of the natural language inference (NLI) task [44] to tabular natural language inference.

This dissertation examines reasoning and inference over semi-structured tabular text, specifically entity-centric InfoBox tables (cf. Figure 1.1). To explore this, we introduce a new dataset called INFOTABS [84] in Chapter 3, which is used to investigate the task of tabular Natural Language Inference (NLI), based on premises that are extracted from Wikipedia info-boxes. INFOTABS's semi-structured, multi-domain and heterogeneous nature of the tabular data admits complex, multi-faceted reasoning. In Chapter 10, we extend the INFOTABS to it's multilingual version XINFOTABS [4, 171], which consist of 10 languages, belonging belong to seven distinct language families (seven continent, 2.76

billion speakers) and six unique writing scripts. To create XINFOTABS, we leverage machine translation models and developed an effective translation pipeline which provide high-quality translations of tabular data.

Tabular Data Scarcity: Human-generated tabular datasets, such as INFOTABS, are limited in scale and thus insufficient for learning with large language models [51, 158]. Since curating these datasets requires expertise, huge annotation time, and expense, they cannot be scaled. Furthermore, it has been shown that these datasets suffer from annotation bias and spurious correlation problem [75, 87, 202]. In contrast, automatically generated data lacks diversity and have naive reasoning aspects. Recently, [173, 190, 306] proposed to use large language generation model [137, 208, 210] for data generation. Despite substantial improvement, these generation approaches still lack factuality, i.e., suffer hallucination, have poor facts coverage, and also suffer from token repetition (refer to Chapter 8 analysis). Recently, [25] shows that automatic tabular NLG frameworks cannot produce logical statements and provide only surface reasoning. To address the above shortcomings, in Chapter 8, we propose a semi-automatic framework that exploits the patterns in tabular structure for hypothesis generation.

1.3 Integrating Knowledge for Tabular Reasoning

Tables hold information in succinct form, which makes information navigation in the cluttered world challenging. Tables lack the necessary context to comprehend the meaning of a text fragment (such as a key) and its relationship to other elements (such as value and other keys). For example, in Figure 1.1, we need to interpret the key *'Length'* in the context of music albums for the given table. Furthermore, due to inadequate training data, models trained on tables are often feeble in implicit lexical knowledge. This affects interpreting the meaning of words such as *"less than"* in H1 (c.f. Figure 1.1). Therefore, in Chapter 4 and Chapter 5, we explore the problem of knowledge integration using simple pre-processing techniques and knowledge graph incorporated transformer long short-term memory (TransKBLSTM) based approaches. We observe that incorporating knowledge not only improves tabular model performance, but also model interpretability. Furthermore, although simple prepossessing is effective, we observe that our transformer LSTM based approach performs better.

1.4 Probing Tabular Reasoning Models

Merely achieving high accuracy is not sufficient evidence of reasoning: the model may arrive at the right answer for the wrong reasons leading to inadequate generalization over unseen data. "Reasoning" is a multi-faceted phenomenon, and fully characterizing it is almost impossible. However, one can probe for the *absence* of evidence-grounded reasoning i.e. "reasoning failures" via model responses to carefully constructed inputs and their variants. For example there are certain pieces of information in the premise (irrelevant to the hypothesis) when changed, should not impact the outcome, thus making the outcome *invariant* to these changes. For example, deleting irrelevant rows from the premise should not change the model's predicted label. Contrary to this is the relevant information ("evidence") in the premise. Changing these pieces of information should vary the outcome in a predictable manner, making the model *covariant* with these changes. For example, deleting relevant evidence rows should change the model's predicted label to NEUTRAL[1].

While "reasoning" can take varied forms, a model that claims to do so should at least ground its outputs on the evidence provided in its inputs. Concretely, we argue that such a model should (a) be self-consistent in its predictions across controlled variants of the input, (b) use the evidence presented to it, and the right parts thereof, and, (c) avoid being biased *against* the given evidence by knowledge encoded in the pre-trained embeddings. Overall, the guiding premise for this (in-/co-)variants perturbation probing is:

> *Any "Evidence-based reasoning" systems should respond predictably to controlled input changes.*

Directly checking for such property there would require a lot of labeled data—a big practical impediment. Fortunately, in the case of tabular semi-structured data, the (in-/co-)variants associated with these dimensions allow controlled and semi-automatic edits to the inputs leading to predictable variation of the expected output. We instantiate the above knowledge along three dimensions to introduce specific probes for experiment on INFOTABS in the dissertation Chapter 6. Our probes demonstrate that models often fail to reason properly on the semi-structured inputs. For example, they often ignore relevant rows, and (a) focus on the irrelevant rows [182], (b) use only the hypothesis sentence [87,

[1]This strategy has been either explicitly or implicitly also employed for recent non-tabular work [72, 221].

202], or (c) knowledge acquired during pre-training [83, 104] . In essence, they use spurious correlations between irrelevant rows, the hypothesis, and the inference label for prediction.

1.5 Evidence Grounded Tabular Inference

When adapted for tabular NLI by flattening tables into synthetic sentences using heuristics, tabular inference model as described in Chapter 3 achieve remarkable performance on the datasets, such as INFOTABS. However, as discussed in tabular probing Chapter 6 [83] models often fail to reason properly on the semi-structured inputs. More specifically they either focus on irrelevant rows [182] or ignore the premise tables [83]. Thus, existing NLI systems optimized solely for the label prediction cannot be trusted. It is not sufficient for a model to be merely *Right* but also *Right for the Right Reasons*. In particular, at least identifying the relevant elements of inputs as the *'Right Reasons'* is essential for trustworthy reasoning. We argue that a reasoning system can be deemed trustworthy only if it exposes how its decisions are made, thus admitting verification of the reasons for its decisions. We address this issue by introducing the task of *Trustworthy Tabular Inference*, in dissertation Chapter 7 where model first extracts relevant rows as evidence and then predict the inference labels. A two-stage sequential prediction approach is proposed for the task, comprising of an evidence extraction stage, followed by an inference stage. In the evidence extraction stage, the model extracts the necessary information needed for the second stage. In the inference stage, the NLI model uses only the extracted evidence as the premise for the label prediction task

1.6 Pre-Training for Enhancing Tabular Reasoning

Recently, [199] shows that LM's pre-trained without explicit supervision on a huge corpus of free web data implicitly incorporate several types of knowledge into their parameters. For extracting this knowledge from language models (LM), various methods utilize probing [95, 265], attention [105, 279], and prompting [200, 237] strategies. This internalized knowledge cannot be retrieved when fine-turning for a subsequent task. One explanation is that the objectives of pre-training and fine-tuning are vastly different. This variation in training objectives also diminishes the expected performance gains of the task, hence necessitating further pre-training on training data [58, 224, 284]. Therefore,

reframing the subsequent task as a joint pre-training objective becomes essential. Hence, in Chapter 9, we reformulate the tabular NLI, i.e., our downstream task as a cloze-style problem, a.k.a, a masked language modeling (MLM) problem. We utilize the efficient Pattern-Exploiting Training (PET) technique [231, 232, 249].

1.7 Dissertation Overview

Thesis Statement: Reasoning on semi-structured data, particularly entity-centric tables, is simple for humans but difficult for NLP models, which are primarily designed for unstructured text. Even when these models appear to make correct inferences, they do so for the wrong reasons. To address these challenges, the models should be able to incorporate knowledge and focus on the relevant parts of semi-structured evidence.

In this dissertation, we explore reasoning and inference over semi-structured tabular text, more precisely entity-centric InfoBox tables, which is a critical component of Natural Language Understanding. The task poses numerous real challenges, including effective table representation, successful knowledge addition, model robustness to perturbation, requisite evidence extraction, addressing tabular data scarcity, and multilingual adaptation. To address these challenges, we introduce novel resources (INFOTABS,XINFOTABS), systematic probes, pattern exploited training, trustworthy modeling framework, and effective data augmentation (AUTO-TNLI) techniques for tables. By tackling these issues in semi-structured data, we hope to contribute to the development of novel methods for reasoning with tabular information, and ultimately advance our understanding of these complex data types.

In this dissertation, we addressed the following research questions:

Q1. How do models designed for unstructured text adapt to (semi-)structured data?

A1. We introduce the task of inference on semi-structure tabular data via INFOTABS datasets and create initial baselines on it. (Chapter 3)

Q2. How does one incorporate knowledge both implicit and explicit type into tabular models?

A2. We study two effective ways to integrate knowledge in tabular reasoning model (a.) simple pre-processing techniques, and (b.) a knowledge transformer LSTM models. (Chapters 4, 5, and 9)

Q3. How to ensure that the model is doing correct evidence-based reasoning?

A3. We design a systematic probes to evaluating tabular models for (a.) robustness to artifacts, (b.) relevant evidence selection, and (c.) robustness to counterfactual changes.

(Chapter 6)

Q4. How to enforce existing model to select right evidence for reasoning?

A4. To address evidence selection issues, we introduce trustworthy tabular inference, a two-stage approach that first extracts evidence and then predicts the inference label. (Chapter 7)

Q5. How to address tabular data scarcity problem for effective data augmentation?

A5. We explore effective semi-automated framework for tabular data enhancement, thus creating AUTO-TNLI for human curated INFOTABS augmentation. (Chapter 8)

Q6. How to effectively pre-trained model for entity-centric semi-structured tables?

A6. We enhance the model's reasoning via prompt learning, i.e., PET, to extracts knowledge from semi-structured tables, to increase model performance, generalizability and robustness, specially on adversarial datasets. (Chapter 9)

Q7. How can we ensure tabular reasoning model reason across multiple language (not just English)?

A7. We extended the English tabular inference dataset (TNLI) INFOTABS to it multilingual variant XINFOTABS, which consists of 10 languages. (Chapter 10)

	Breakfast in America	Relevance
Released[4]	29 March 1979[4]	H3
Recorded[3,4]	May-December 1978[3,4]	H2, H3
Studio	The Village Recorder in Los Angeles[3]	
Genre	Pop, Art Rock, Soft Rock	
Length[2]	46:06[2]	H1
Label	A&M	
Producer[1]	Peter Henderson, Supertramp[1]	H1

H1: Supertramp produced[1] an album that was less than an hour long[2].

H2: Most of Breakfast in America was recorded[3] in the last month of 1978[3].

H3: Breakfast in America was released[4] the same month recording[4] ended.

Figure 1.1: A semi-structured premise (the table 'Breakfast in America') example from InfoTabS. The table displays three hypotheses, with H1 entailed, H2 neither entailed nor contradictory, and H3 contradictory. Relevant rows are highlighted in color, and the "Relevance" column indicates which hypotheses use each row for reasoning.

CHAPTER 2

BACKGROUND

2.1 Natural Language Inference

Textual Entailment (TE) and Natural Langauge Inference (NLI) are two related tasks in Natural Language Processing (NLP) that involve determining the logical relationship between two pieces of text. Both tasks involves determining whether a given hypothesis is true (or more generally, whether it is entailed, contradicted, or neutral) given a certain context or premise. Now consider the following pair of sentences:

Premise: The cat chased the mouse. Hypothesis: The cat ate the mouse.

In this case, the hypothesis (The cat ate the mouse) is neither entailed nor contradicted by the premise (the cat chased the mouse), because it is possible that the cat ate the mouse after the cat chased, or maybe not if the mouse succeeded in escaping. Therefore, the relationship between the premise and the hypothesis is considered ”neutral” i.e. maybe true or false in NLI.

Here is few more examples:

Premise: John bought a new bike yesterday. Hypothesis: John has a bike today.

In this case, the premise (John bought a new bike yesterday) entails the hypothesis (John has a bike today), because if John bought a bike yesterday, then it is likely that he still has it today.

Premise: The restaurant was completely full and had a long waiting list. Hypothesis: The food was delicious.

In this case, the hypothesis (the food was delicious) is not entailed by the premise (the restaurant was completely full and had a long waiting list), because it is possible that people were waiting for a table despite the food being mediocre or bad. Therefore, the relationship between the premise and the hypothesis is considered ”neutral” i.e. maybe true or false in NLI.

Both, NLE and TE are well studied in the past with several diverse datasets. The annual PASCAL RTE challenges [44] were associated with several thousands of human-annotated entailment pairs. The Stanford Natural Language Inference (SNLI) dataset [17] is the first large scale entailment dataset that uses image captions as premises, while the Multi-Genre Natural Language Inference (MNLI) [280] uses premises from multiple domains. The QNLI and WNLI datasets provide a new perspective by converting the SQuAD question answering data [214] and Winograd Schema Challenge data [136] respectively into inference tasks. More recently, SciTail [124] and Adversarial NLI [185] have focused on building adversarial datasets; the former uses information retrieval to select adversarial premises, while the latter uses iterative annotation cycles to confuse models. NLI and TE are important tasks in NLP because they have many practical applications, such as question answering, summarization, and dialogue systems, among others. They also serve as a benchmark for evaluating NLP models performance.

2.2 Reasoning for Natural Language Inference

Let's consider two examples to understand different types of reasoning require for NLI:

Premise: A woman is slicing an onion. Hypothesis: The woman is making a salad.

To correctly classify this example, the model needs to understand the relationship between slicing an onion and making a salad. The model must also understand that the given hypothesis is consistent with the given premise.

Premise: The company announced a new product today. Hypothesis: A product launch occurred today.

In this example, the model needs to understand that "announcing a new product" and "product launch" are semantically similar, and that the given hypothesis is consistent with the given premise. The model must also recognize that the genre of the premise and hypothesis may differ, and this should not affect the classification. To perform NLI effectively, a model needs to be able to perform several types of reasoning, some including:

1. **Semantic understanding:** This involves understanding the meaning of words and phrases in the premise and hypothesis sentences, and how they relate to each other. This requires the model to be able to recognize word-level and sentence-level semantics, such as synonyms, antonyms, hyponyms, and hypernyms and paraphrases. For

example, in the first example ”A woman is slicing an onion” and ”The woman is making a salad,” the model must recognize that ”slicing an onion” and ”making a salad” are related concepts, and that they both involve food preparation.

2. **Logical reasoning:** This involves the ability to perform logical reasoning, such as recognizing contradictions and entailment relationships between the premise and hypothesis sentences. The model must be able to recognize when the hypothesis contradicts the premise, when it is entailed by the premise, or when it is neutral with respect to the premise. For example, in the first example, the model must recognize that ”slicing an onion” does not necessarily entail ”making a salad,” but that it is consistent with it.
3. **World knowledge:** This involves having access to a broad range of world knowledge, including common sense and domain-specific knowledge, to make accurate predictions. The model must be able to recognize when a hypothesis is consistent with the real world, and when it is not. For example, in the second example ”The company announced a new product today” and ”A product launch occurred today,” the model must recognize that the two statements are referring to the same event, even though they use different phrasing.
4. **Contextual reasoning:** This involves considering the context in which the premise and hypothesis sentences appear, including the domain, genre, style, and discourse structure, to make accurate predictions. The model must be able to recognize when the context of the premise and hypothesis sentences differ, and when this should not affect the inference. For example, in the second example, the model must recognize that the domain of the premise sentence is a news article, while the domain of the hypothesis sentence is more general.

Across NLP, a lot of work has been published around different kinds of reasonings. The GLUE [270] benchmark introduced collection of tasks (NLI, QA, Textual Similarity and soon) and datasets (MNLI, RTE, QNLI and soon) along with various reasoning types (Numerical, Temporal, Ellipsys, Coreference, Entity Typing, Knowledge and Common Sense , Quantification and soon), required to evaluate the performance of natural language understanding models on the diverse range of NLP tasks, including NLI. Recently, SuperGlue [269] benchmark is also introduced which introduce more diverse task and

even more complex reasoning. Challenging datasets have emerged in NLI that emphasize distinct complex reasoning. [14] pose the task of determining the most plausible inferences based on observation, requiring abductive reasoning. Others such as, common sense reasoning in [247], temporal reasoning in [308], numerical reasoning focused in [178, 267] and multi-hop [122] reasoning have all sparked immense research interest.

2.3 Table Natural Language Inference

Recently NLP community has focused on investigating various NLP tasks on diverse types of semi-structured tabular data (refer to Figure 2.1 for table types), including tabular NLI and fact verification [26, 5 4]. As of 2023, other than the work presented in this dissertation, there is only one public human curated NLI dataset on tables, namely TabFact [26]. In this section you will focus on TabFact, and will present our new dataset INFOTABS on entity-centric tables (refer to Chapter 3). TabFact [26], considers database-style tables as premises with human-annotated hypotheses sentence for inference task. The Wikipedia tables of TabFact are homogeneous, with each column having structural redundancy and common entity type. Figure 1.1 and Figure 2.2 show inference examples for INFOTABS and TabFact datasets respectively. The reasonings of the hypotheses in TabFact are numerical logical operations based mostly involving comparatives, superlative, counting, ranking, aggregation and soon, refer to Figure 2.3. TabFact based on complexity of reasoning has separate simple and complex test sets.

Manual verses Automated Dataset: Tabular dataset has long been explored [119, 177, 225, 284]. For tabular NLI in particular, the datasets can be categorized into (1.) Manually created datasets [84] (Chapter 3) with manually creates both hypothesis and premise, [26] manually creates the hypothesis while premise is automatically generated (2.) Synthetically created semi-automatically generated datasets which completely automate data generation requires manual designing table-dependent context-free grammar (CFG) [58], or require logical forms to be annotated [25, 29, 177]. These work mostly address the database style simialr to TabFact types tables. Furthermore, semi-automatic systems requiring a Context Free Grammar (CFG) or logical forms contains reasoning which is often limited to certain types. Creating sentences that contain other reasonings (like lexical reasoning, knowledge, and common sense reasoning) is challenging using CFG and logical forms. In

this dissertation Chapter 8, we introduce AUTO-TNLI a semi-automatically synthetic and counterfactual INFOTABS style large scale dataset.

Other Tabular Tasks: Additionally, various question answering and semantic parsing tasks [1, 24, 27, 128, 151, 189, 195, 245, 297, 302, 303], and table-to-text generation [25, 141, 194, 207, 293] are also recently introduce. Previous work has also touched upon semantic parsing and question answering [123, 195], which typically work with tables with many entries that resemble database records.

Table Modeling Work: Some recent papers have also proposed ideas for representing Wikipedia relational tables, some such papers are TAPAS [94], StrucBERT [257], Table2vec [301], TaBERT [291], TABBIE [101], TabStruc [300], TabGCN [204], RCI [77], TURL [49] and TableFormer [289]. Some papers such as [58, 177, 182, 238, 294, 295] study the improvement of tabular inference by pre-training. In this dissertation Chapter 9, we also introduce pattern exploited pre-training (PET) for infobox style tables for INFOTABS dataset.

2.4 Information Extraction for Semi-Structured Tables

There are several approaches for information extraction (IE) from semi-structured text [55]. Below we describe some of them based on amount of data supervision:

1. **Supervised Wrapper Induction:** This is a closed supervised IE (i.e.) approach which infers rules for each relation schema in semi-structure database. These learnt rules are called wrappers. The main idea is to learn rules that are resilient to small page alterations by using locally consistent features surrounding an attribute's values. Prior works includes [20, 43, 67, 81, 131, 132], and others. The main issue with these approaches is that they rely on correct manually labeled data, which limits their scalability. Furthermore, these models rely on specific templates and cannot generalize to other templates.
2. **Distant Supervision Approaches:** These closed IE approaches use distant(auxiliary) supervision to generate cheap, but noisy training data. For examples, using seed knowledge based of one domain as source as distant supervision for other similar domain. Prior work includes [36, 66, 90, 159], and others. One significant disadvantage of such methods is that they necessitate the availability of a domain-specific knowledge base for each domain. Furthermore, models have low recall score due to closed form nature of task.

3. **Open IE - Schema Less Approaches:** These methods are designed to extract new and unknown relations from semi-structured text. The essential aspect is to take advantage of data redundancy in data-rich websites that overlap at the schema and instance level. The methods locate extractors that maximize overlap semantically comparable data from multiple sources [18]. Recent methods such as [160, 161] discover new connections using visual similarity between seed and new (relation, object) pairings. The disadvantage of these methods is that they require numerous websites within a domain for data redundancy.

Apart from the approaches outlined above, two shared tasks, namely the SemEval'21 Task 9 [226] and FEVEROUS'21 shared task [5] are also proposed in the past, both of which include IE for relational tables. The two IE task are: (a.) **SemEval Task 9:** This task focuses on statement verification and evidence extraction using relational tables derived from scientific articles. Some methods proposed for these shared tasks are [2, 74, 113, 177, 227, 261], and (b.) **FEVEROUS'21:** The FEVEROUS'21 shared task verifies information using unstructured and structured evidence from open-domain Wikipedia. FEVEROUS data has relational tables, unstructured text, and entity tables. Some methods proposed for these shared tasks are [16, 65, 76, 127, 165, 229, 255].

2.4.1 Web-Table Extraction

The majority of the approaches presented above are intended for semi-structured pages. Unique web-table extraction methods are provided for tables in INFOTABS and TabFact. These methods intend to extract semantics from tables by determining the subject column, column class, and ontological relations for pairs of columns. The subject column methods [100, 264] utilize generic subject entity characteristics such as value uniqueness, string type, amount of characters, and words for IE. The column class methods [48, 271] utilize external data – web extracted triples, knowledge graph for IE. The relational pair methods [82, 148, 264] utilize similarity measure between a column and entities of a type in a knowledge base for IE.

2.5 Multilinguality and Other Concerns

Given the need for greater inclusivity towards linguistic diversity in NLP applications, various multilingual versions of datasets have been created for text classification

[42, 203, 290], question answering [8, 37, 138] and structure prediction [188, 211]. Following the introduction of datasets, multilingual leaderboards like XTREME leaderboard [98], the XGLUE leaderboard [147] and the XTREME-R leaderboard [228] have been created to test models' cross-lingual transfer and language understanding.

Multilingual models can be broadly classified into two variants: (a) Natural Language Understanding (NLU) models like mBERT [51], XLM [41], XLM-R [40], XLM-E [32], Rem-BERT [35], and (b) Natural Language Generation (NLG) models like mT5 [286], mBART [157], M2M100 [59]. NLU models have been used in multilingual language understanding tasks like sentiment analysis, semantic similarity and natural language inference while NLG models are used in generation tasks like question-answering and machine translation. Furthermore, multilingual models have shown to be extremely memory and time efficient. Various models have been proposed that achieve state-of-the-art results on the previously mentioned leaderboards.

Multilingual, and specifically cross-Lingual transfer [50, 197], has been widely discussed in the context of low resource languages. Several datasets [8, 42, 139, 188, 203, 290], benchmarks and leaderboards [98, 121, 146, 228], and evaluation frameworks [117, 244, 254] have emerged which focus entirely on evaluation of multilingual NLU. Further, multilingual language models have been developed for (a.) Natural Language Understanding [32, 35, 40, 41, 51], (b.) and Natural Language Generation [59, 286]. Multilingual models have shown great success in cross-lingual transfer and understanding for both languages with varying resource levels.

The Annotation Artifacts Problem: Recently, pre-trained transformer-based models [209] have seemingly outperformed human performance on several NLI tasks in Glue and SuperGlue. However, it has been shown by [78, 87, 179, 187, 202, 266] that these models exploit spurious patterns (artifacts) in the data to obtain good performance. It is imperative to produce datasets that allow for controlled study of artifacts. A popular strategy today is to use adversarial annotation [185, 298] and rewriting of the input [26]. In this dissertation, we shows that one can systematically construct tabular test sets for studying artifacts along specific dimensions.

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

(a.) Tabfact: Database Style Tables

New York Stock Exchange	
Type	Stock exchange
Location	New York City, New York, U.S.
Founded	May 17, 1792; 226 years ago
Currency	United States dollar
No. of listings	2,400
Volume	US$20.161 trillion (2011)

(b.) InfoTabS: Entity Centric (InfoBox) Tables

Boxe (fr)	
Focus	Punching, frappe
Sport olympique	688 av. J.-C. (Grèce ancienne), 1904 (moderne)
Parentalité	Bare-knuckle boxe
Pays d'origine	Préhistorique
Aussi connu sous le nom	Western Boxing, Pugilism Voir note.

(c.) XInfoTabS: Entity Centric Multilingual Tables

Figure 2.1: Types of tables in various tabular inference datasets.

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Entailed Statement

1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statement

1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.

Figure 2.2: Example from TabFact dataset.

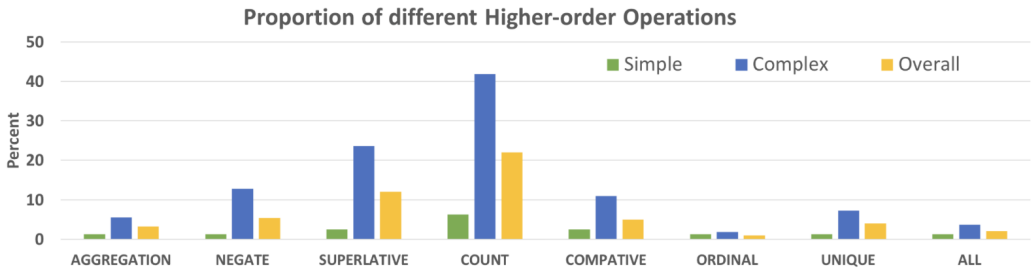


Figure 2.3: Reasoning and their proportion in TabFact dataset.

CHAPTER 3

INFERENCE ON SEMI-STRUCTURED TABLES

Adapted from V. Gupta, M. Mehta, P. Nokhiz, and V. Srikumar, *INFOTABS: Inference on tables as semi-structured data*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 5–10, 2020, Association for Computational Linguistics, pp. 2309–2324.

In this chapter, we introduce the INFOTABS dataset to study and model inference with semi-structured data. Premises in our dataset consist of info-boxes that convey information implicitly, and thus require complex reasoning to ascertain the validity of hypotheses. For example, determining that the hypothesis H2 in Figure 3.1 entails the premise table requires looking at multiple rows of the table, understanding the meaning of the row labeled *Mixed gender*, and also that *Dressage* is a sport.

INFOTABS consists of 23,738 premise-hypothesis pairs, where all premises are info-boxes, and the hypotheses are short sentences. As in the NLI task, the objective is to ascertain whether the premise entails, contradicts or is unrelated to the hypothesis. The dataset has 2,540 unique info-boxes drawn from Wikipedia articles across various categories, and all the hypotheses are written by Amazon's Mechanical Turk workers. Our analysis of the data shows that ascertaining the label typically requires the composing of multiple types of inferences across multiple rows from the tables in the context of world knowledge. Separate verification experiments on subsamples of the data also confirm the high quality of the dataset.

We envision our dataset as a challenging testbed for studying how models can reason about semi-structured information. To control for the possibility of models memorizing superficial similarities in the data to achieve high performance, in addition to the standard train/dev/test split, our dataset includes two additional test sets that are constructed by systematically changing the surface forms of the hypothesis and the domains of the tables.

We report the results of several families of approaches representing word overlap based models, models that exploit the structural aspect of the premise, and also derivatives of state-of-the-art NLI systems. Our experiments reveal that all these approaches underperform across the three test sets. This work is published at ACL 2020 as [84].[1]

3.1 Contributions

The main contributions we make here are:

1. We propose a new English natural language inference dataset, INFOTABS, to study the problem of reasoning about semi-structured data.
2. To differentiate models' ability to reason about the premises from their memorization of spurious patterns, we created three challenge test sets with controlled differences that employ similar reasoning as the training set.
3. We show that several existing approaches for NLI underperform on our dataset, suggesting the need for new modeling strategies.

3.2 Background

Tasks based on semi-structured data in the form of tables, graphs and databases (with entries as text) contain complex reasoning [26, 54] has been studied before. Previous work has also touched upon semantic parsing and question answering [123, 195, and references therein], which typically work with tables with many entries that resemble database records.

Our work is most closely related to TabFact [26], which considers database-style tables as premises with human-annotated hypotheses to form an inference task. While there are similarities in the task formulation scheme, our work presents an orthogonal perspective: (i) The Wikipedia tables premises of TabFact are homogeneous, i.e., each column in a table has structural redundancy and all entries have the same type. One can look at multiple entries of a column to infer extra information, e.g., all entries of a column are about locations. On the contrary, the premises in our dataset are heterogeneous. (ii) TabFact only considers entailment and contradiction; we argue that inference is non-binary with a third "undetermined" class (neutrals). (iii) Compared to our multi-faceted reasonings, the

[1]The dataset, along with associated scripts, are available at https://infotabs.github.io/.

reasonings of the hypotheses in TabFact are limited and mostly numerical or comparatives. (iv) The α_2 and α_3 sets help us check for annotation and domain artifacts.

3.3 Case for Reasoning

We often encounter textual information that is neither unstructured (i.e., raw text) nor strictly structured (e.g., databases). Such data, where a structured scaffolding is populated with free-form text, can range from the highly verbose (e.g., web pages) to the highly terse (e.g. fact sheets, information tables, technical specifications, material safety sheets). Unlike databases, such semi-structured data can be heterogeneous in nature, and not characterized by pre-defined schemas. Moreover, we may not always have accompanying explanatory text that provides context. Yet, we routinely make inferences about such heterogeneous, incomplete information and fill in gaps in the available information using our expectations about relationships between the elements in the data.

Understanding semi-structured information requires a broad spectrum of reasoning capabilities. We need to understand information in an ad hoc layout constructed with elements (cells in a table) that are text snippets, form fields or are themselves sub-structured (e.g., with a list of elements). Querying such data can require various kinds of inferences. At the level of individual cells, these include simple lookup (e.g., knowing that *dressage takes place in an arena*), to lexical inferences (e.g., understanding that *Mixed Gender* means both men and women compete), to understanding types of text in the cells (e.g., knowing that the number 1912 is a year). Moreover, we may also need to aggregate information across multiple rows (e.g., knowing that *dressage is a non-contact sport that both men and women compete in*), or perform complex reasoning that combines temporal information with world knowledge.

We argue that a true test of reasoning should evaluate the ability to handle such semi-structured information. To this end, we define a new task modeled along the lines of NLI, but with tabular premises and textual hypotheses, and introduce a new dataset INFOTABS for this task.

3.4 The Need for Multi-Faceted Evaluation

Before describing the new dataset, we will characterize our approach for a successful evaluation of automated reasoning.

Recent work has shown that many datasets for NLI contain annotation biases or artifacts [202]. In other words, large models trained on such datasets are prone to learning spurious patterns—they can predict correct labels even with incomplete or noisy inputs. For instance, *not* and *no* in a hypothesis are correlated with contradictions [187]. Indeed, classifiers trained on the hypotheses only (ignoring the premises completely) report high accuracy; they exhibit *hypothesis bias*, and achieving a high predictive performance does not need models to discover relationships between the premise and the hypothesis. Other artifacts are also possible. For example, annotators who generate text may use systematic patterns that "leak" information about the label to a model. Or, perhaps models can learn correlations that mimic reasoning, but only for one domain. With millions of parameters, modern neural networks are prone to overfitting to such imperceptible patterns in the data.

From this perspective, if we seek to measure a model's capability to understand and reason about inputs, we cannot rely on a single fixed test set to rank models. Instead, we need multiple test sets (of similar sizes) that have controlled differences from each other to understand how models handle changes along those dimensions. While all the test sets address the same task, they may not all be superficially similar to the training data.

With this objective, we build three test sets, named α_1, α_2 and α_3. Here, we briefly introduce them; §3.5 goes into specifics. Our first test set (α_1) has a similar distribution as the training data in terms of lexical makeup of the hypotheses and the premise domains.

The second, *adversarial test set (α_2)*, consists of examples that are also similar in distribution to the training set, but the hypothesis labels are changed by expert annotators changing as few words in the sentence as possible. For instance, if *Album X was released in the* 21^{st} *century* is an entailment, the sentence *Album X was released before the* 21^{st} *century* is a contradiction, with only one change. Models that merely learn superficial textual artifacts will get confused by the new sentences. For α_2, we rewrite entailments as contradictions and vice versa, while the neutrals are left unaltered.

Our third test set is the *cross-domain (α_3)* set, which uses premises from domains that are not in the training split, but generally, necessitate similar types of reasoning to arrive at

the entailment decision. Models that overfit domain-specific artifacts will underperform on α_3.

Note that, in this work, we describe and introduce three different test sets, but we expect that future work can identify additional dimensions along which models overfit their training data and construct the corresponding test sets.

3.5 The InfoTabS Dataset

In this section, we will see the details of the construction of INFOTABS. We adapted the general workflow of previous crowd sourcing approaches for creating NLI tasks [17] that use Amazon's Mechanical Turk.

3.5.1 Sources of Tables

Our dataset is based on 2,540 unique info-boxes from Wikipedia articles across multiple categories (listed in Section 3.5.4). We did not include tables that have fewer than 3 rows, or have non-English cells (e.g., Latin names of plants) and technical information that may require expertise to understand (e.g., astronomical details about exoplanets). We also removed non-textual information from the table, such as images. Finally, we simplified large tables into smaller ones by splitting them at sub-headings. Our tables are isomorphic to key-value pairs, e.g., in Figure 3.1, the bold entries are the keys, and the corresponding entries in the same row are their respective values.

3.5.2 Sentence Generation

Annotators were presented with a tabular premise and instructed to write three self-contained grammatical sentences based on the tables: one of which is true given the table, one which is false, and one which may or may not be true. The turker instructions included illustrative examples using a table and also general principles to bear in mind, such as avoiding information that is not widely known, and avoiding using information that is not in the table (including names of people or places). The turkers were encouraged not to restate information in the table, or make trivial changes such as the addition of words like *not* or changing numerical values. We refer the reader to the project website for a snapshot of the interface used for turking, which includes the details of instructions.

We restricted the turkers to be from English-speaking countries with at least a Master's

qualification. We priced each HIT (consisting of one table) at 50¢. Following the initial turking phase, we removed grammatically bad sentences and rewarded workers whose sentences involved multiple rows in the table with a 10% bonus.

3.5.3 Data Partitions

We annotated 2,340 unique tables with nine sentences per table (i.e., three turkers per table).[2] We partitioned these tables into training, development (Dev), α_1 and α_2 test sets. To prevent an outsize impact of influential turkers in a split, we ensured that the annotator distributions in the Dev and test splits are similar to that of the training split.

We created the α_2 test set from hypotheses similar to those in α_1, but from a separate set of tables, and perturbing them as described in §3.4. On an average, $\sim$ 2.2 words were changed per sentence to create α_2, with no more than 2 words changing in 72% of the hypotheses. The provenance of α_2 ensures that the kinds of reasoning needed for α_2 are similar to those in α_1 and the development set. For the α_3 test set, we annotated 200 additional tables belonging to domains not seen in the training set (e.g., diseases, festivals). As we will see in §3.8, hypotheses in these categories involve a set of similar types of reasonings as α_1, but with different distributions.

In total, we collected 23,738 sentences split almost equally among entailments, contradictions, and neutrals. Table 3.1 shows the number of tables and premise-hypothesis pairs in each split. In all the splits, the average length of the hypotheses is similar. We refer the reader to next section for additional statistics about the data.

3.5.4 INFOTABS Dataset Statistics

In this section, we provide some essential statistics that will help in a better understanding of the dataset.

Table 3.2 shows a split-wise analysis of premises and annotators. The table shows that there is a huge overlap between the train set and the other splits except α_3. This is expected since α_3 is from a different domain. Also, we observe that tables in α_3 are longer. In the case of annotators, we see that most of our dataset across all splits was annotated by the

[2]For tables with ungrammatical sentences, we repeated the HIT. As a result, a few tables in the final data release have more than 9 hypotheses.

same set of annotators.

Table 3.3 presents information on the generated hypotheses. The table lists the average number of words in the hypotheses. This is important because a dissimilar mean value of words would induce the possibility of length bias, i.e., the length of the sentences would be a strong indicator for classification.

Table 3.4 shows the overlap between hypotheses and premise tables across various splits. Stop words like *a, the, it, of,* etc. are removed. We observe that the overlap is almost similar across labels.

Tables 3.5 and 3.6 show the distribution of table categories in each split. We accumulate all the categories occurring for less than 3% for every split into the "Other" category.

3.5.5 Validating Hypothesis Quality

We validated the quality of the data using Mechanical Turk. For each premise-hypothesis in the development and the test sets, we asked turkers to predict whether the hypothesis is entailed or contradicted by, or is unrelated to the premise table. We priced this task at 36¢ for nine labels.

The inter-annotator agreement statistics are shown in Table 3.7, with detailed statistics in Section 3.6. On all splits, we observed significant inter-annotator agreement scores with Cohen's Kappa scores [9] between 0.75 and 0.80. In addition, we see a majority agreement (at least 3 out of 5 annotators agree) of range between 93% and 97%. Furthermore, the human accuracy agreement between the majority and gold label (i.e., the label intended by the writer of the hypothesis), for all splits is in range 80% to 84%, as expected given the difficulty of the task.

3.6 Statistics of INFOTABS Verification

Table 3.8 shows the detailed agreement statistics of verification for the development and the three test splits. For every premise-hypothesis pair, we asked five annotators to verify the label. The table details the verification agreement among the annotators, and also reports how many of these majority labels match the gold label (i.e., the label intended by the author of the hypothesis). We also report individual annotator label agreement by matching the annotator's label with the gold label and majority label for an example.

Finally, the table reports the Fleiss Kappa (across all five annotation labels) and the Cohen Kappa (between majority and gold label) for the development and the three test splits.

We see that, on average, about 84.8% of individual labels match with the majority label across all verified splits. Also, an average of 75.15% individual annotations also match the gold label across all verified splits.

From Table 3.8, we can calculate the percentage of examples with at least 3, 4, and 5 label agreements across 5 verifiers for all splits. For all splits, we have very high inter-annotator agreement of >95.85% for at-least 3, > 74.50% for at-least 4 and 43.91% for at-least 5 annotators. The number of these agreements match with the gold label are: >81.76% for at-least 3, > 67.09% for at-least 4 and 40.85% for at-least 5 for all splits.

3.7 Reasoning in INFOTABS

Our inventory of reasoning types is based on GLUE diagnostics [270], but is specialized to the problem of reasoning about tables. Consequently, some categories from GLUE diagnostics may not be represented here, or may be merged into one category.

We assume that the table is correct and complete. The former is always true for textual entailment, where we assume that the premise is correct. The latter need not be generally true. However, in our analysis, we assume that the table lists all the relevant information for a field. For example, in a table for a music group as in Figure 3.2, if there is a row called **Labels**, we will assume that the labels listed in that row are the only labels associated with the group.

Note that a single premise-hypothesis pair may be associated with multiple types of reasoning. If the same reasoning type is employed multiple times in the same pair, we only mark it once. All definitions and their boundaries were verified via several rounds of discussions. Following this, three graduate students independently annotated 160 pairs from the Dev and α_3 test sets each, and edge cases were adjudicated to arrive at consensus labels.

3.7.1 Simple Lookup

This is the simple case where there is no reasoning, and the hypothesis is formed by literally restating information in the table. For example, using the table in Figure 3.3,

Femme aux Bras Croisés is privately held. is a simple lookup.

3.7.2 Multi-Row Reasoning

Multiple rows in the table are needed to make an inference. This has the strong requirement that without multiple rows, there is no way to arrive at the conclusion. Exclude instances where multiple rows are used only to identify the type of the entity, which is then used to make an inference. The test for multi-row reasoning is: If a row is removed from the table, then the label for the hypothesis may change.

3.7.3 Entity Type

Involves ascertaining the type of an entity in question (perhaps using multiple rows from the table), and then using this information to make an inference about the entity.

This is separate from multi-row reasoning even if discovering the entity type might require reading multiple rows in the table. The difference is a practical one: we want to identify how many inferences in the data require multiple rows (both keys and values) separately from the ones that just use information about the entity type. We need to be able to identify an entity and its type separately to decide on this category. In addition, while multi-row reasoning, by definition, needs multiple rows, entity type may be determined by looking at one row. For instance, looking at Figure 3.3, one can infer that the entity type is a *painting* by only looking at the row with key value **Medium**. Lastly, ascertaining the entity type may require knowledge, but if so, then we will not explicitly mark the instance as Knowledge & Common Sense. For example, knowing that SNL is a TV show will be entity type and not Knowledge & Common Sense.

3.7.4 Lexical Reasoning

Any inference that can be made using words, independent of the context of the words falls. For example, knowing that dogs are animals, and alive contradicts dead would fall into the category of lexical reasoning. This type of reasoning includes substituting words with their synonyms, hypernyms, hyponyms and antonyms. It also includes cases where a semantically equivalent or contradicting word (perhaps belonging to a different root word) is used in the hypothesis., e.g., replacing understand with miscomprehend. Lexical reasoning also includes reasoning about monotonicity of phrases.

3.7.5 Negation

Any explicit negation, including morphological negation (e.g., the word *affected* being mapped to *unaffected*). Negation changes the morphology without changing the root word, e.g., we have to add an explicit *not*.

This category includes double negations, which we believe is rare in our data. For example, the introduction of the phrase *not impossible* would count as a double negation. If the word *understand* in the premise is replaced with *not comprehend*, we are changing the root word (understand to comprehend) and introducing a negation. So this change will be marked as both Lexical reasoning and Negation.

3.7.6 Knowledge and Common Sense

This category is related to the World Knowledge and Common Sense categories from GLUE. To quote the description from GLUE: "... the entailment rests not only on correct disambiguation of the sentences, but also application of extra knowledge, whether it is concrete knowledge about world affairs or more common-sense knowledge about word meanings or social or physical dynamics."

While GLUE differentiates between world knowledge and common sense, we found that this distinction is not always clear when reasoning about tables. So we do not make the distinction.

3.7.7 Named Entities

This category is identical to the Named Entities category from GLUE. It includes an understanding of the compositional aspect of names (for example, knowing that the *University of Hogwarts* is the same as Hogwarts). Acronyms and their expansions fall into this category (e.g., the equivalence of *New York Stock Exchange* as *NYSE*).

3.7.8 Numerical Reasoning

Any form of reasoning that involves understanding numbers, counting, ranking, intervals and units falls under this group. This category also includes numerical comparisons and the use of mathematical operators to arrive at the hypothesis.

3.7.9 Temporal Reasoning

Any inferences that involves reasoning about time fall into this category. There may be an overlap between other categories and this one. Any numerical reasoning about temporal quantities and the use of knowledge about time should be included here. Examples of temporal reasoning: (a.) 9 AM is in the morning. (Since this is knowledge about time, we will only tag this as Temporal.), (b.) 1950 is the 20^{th} century., (c.) 1950 to 1962 is twelve years., and (d.) Steven Spielberg was born in the winter of 1946. (If the table has the date—18th December, 1946—and the location of birth—Ohio, this sentence will have both knowledge & Common Sense and temporal reasoning. This is because one should be able to tell that the birth location is in the northern hemisphere (knowledge) and December is part of the Winter in the northern hemisphere (temporal reasoning)).

3.7.10 Coreference

This category includes cases where expressions refer to the same entity. However, we do not include the standard gamut of coreference phenomena in this category because the premise is not textual. We specifically include the following phenomena in this category: Pronoun coreference, where the pronoun in a hypothesis refers to a noun phrase either in the hypothesis or the table. E.g., *Chris Jericho lives in a different state than he was born in.* A noun phrase (not a named entity) in the hypothesis refers to a name of an entity in the table. For example, the table may say that *Bob has three children, including John* and the hypothesis says that *Bob has a son*. Here the phrase *a son* refers to the name *John*.

If there is a pronoun involved, we should not treat it as entity type or knowledge even though knowledge may be needed to know that, say, *Theresa May* is a woman and so we should use the pronoun *she*.

To avoid annotator confusion, when two names refer to each other, we label it only as the Named Entities category. For example, if the table talks about *William Henry Gates III* and the hypothesis describes *Bill Gates*, even though the two phrases do refer to each other, we will label this as Named Entities.

3.7.11 Quantification

Any reasoning that involves introducing a quantifier such as every, most, many, some, none, at least, at most, etc. in the hypothesis. This category also includes cases where

prefixes such as multi- (e.g., *multi-ethnic*) are used to summarize multiple elements in the table.

To avoid annotator confusion, we decide that the mere use of quantifiers like most and many is quantification. However, if the quantifier is added after comparing two numerical values in the table, the sentence is labeled to have numerical reasoning as well.

3.7.12 Subjective/Out of Table

Subjective inferences refer to any inferences that involve either value judgment about a proposition or a qualitative analysis of a numerical quantity. Out of table inferences involve hypotheses that use extra knowledge that is neither a well known universal fact nor common sense. Such hypotheses may be written as factive or implicative constructions. Below are some examples of this category: (a.) Based on a table about Chennai: *Chennai is a very good city.*, (b.) If the table says that John's height is 6 feet, then the hypothesis that *John is a tall person.* may be subjective. However, if John's height is 8 feet tall, then the statement that *John is tall.* is no longer subjective, but common sense., (c.)If the table only says that John lived in Madrid and Brussels, and the hypothesis is *John lived longer in Madrid than Brussels.* This inference involves information that is neither well known nor common sense., and (d.) Based on the table of the movie Jaws, the hypothesis *It is known that Spielberg directed Jaws* falls in this category. The table may contain the information that Spielberg was the director, but this may or may not be well known. The latter information is out of the table.

3.7.13 Syntactic Alternations

This refers to a catch-all category of syntactic changes to phrases. This includes changing the preposition in a PP, active-passive alternations, dative alternations, etc. We expect that this category is rare because the premise is not text. However, since there are some textual elements in the tables, the hypothesis could paraphrase them.

This category is different from reasoning about named entities. If a syntactic alternation is applied to a named entity (e.g., *The Baltimore City Police* being written as *The Police of Baltimore City*), we will label it as a Named Entity if, and only if, we consider both phrases as named entities. Otherwise, it is just a syntactic alternation. Below are some examples of this category: (a.) *New Orleans police officer* being written as *police officer of New Orleans*.,

and (b.) *Shakespeare's sonnet* being written as *sonnet of Shakespeare*.

3.7.14 Ellipsis

This category is similar in spirit to the category Ellipsis/Implicits in GLUE: "An argument of a verb or another predicate is elided in the text, with the reader filling in the gap." Since in our case, the only well-formed text is in the hypothesis, we expect such gaps only in the hypothesis. (Compared to GLUE, where the description makes it clear that the gaps are in the premises and the hypotheses are constructed by filling i n t he g aps w ith either correct or incorrect referents.). For example, in a table about Norway that lists the per capita income as $74K, the hypothesis that *The per capita income is $74K.* elides the fact that this is about citizens of Norway, and not in general.

3.8 Reasoning Analysis

Figures 3.4 and 3.5 summarize these annotation efforts. We see that we have a multi-faceted complex range of reasoning types across both sets. Importantly, we observe only a small number of simple lookups, simple negations for contradictions, and mere syntactic alternations that can be resolved without complex reasoning. Many instances call for looking up multiple rows, and involve temporal and numerical reasoning. Indeed, as Figures 3.6 and 3.7 show a large number of examples need at least two distinct kinds of reasoning; on an average, sentences in the Dev and α_3 sets needed 2.32 and 1.79 different kinds of reasoning, respectively.

We observe that semi-structured premises forced annotators to call upon world knowledge and common sense (KCS); 48.75% instances in the Dev set require KCS. (In comparison, in the MultiNLI data, KCS is needed in 25.72% of examples.) We conjecture that this is because information about the entities and their types is not explicitly stated in tables, and have to be inferred. To do so, our annotators relied on their knowledge about the world including information about weather, seasons, and widely known social and cultural norms and facts. An example of such common sense is the hypothesis that *"X was born in summer"* for a person whose date of birth is in May in New York. We expect that the INFOTABS data can serve as a basis for studying common sense reasoning alongside other recent work such as that of [247].

Neutral hypotheses are more inclined to being subjective/out-of-table because almost anything subjective or not mentioned in the table is a neutral statement. Despite this, we found that in all evaluations in Section 3.9.3.6 (except those involving the adversarial α_2 test set), our models found neutrals almost as hard as the other two labels, with only an $\approx$ 3% gap between the F-scores of the neutral label and the next best label.

The distribution of train, dev, α_1 and α_2 are similar because the premises are taken from the same categories. However, tables for α_3 are from different domains, hence not of the same distribution as the previous splits. This difference is also reflected in Figures 3.4 and 3.5, as we see a different distribution of reasonings for each test set. This is expected; for instance, we cannot expect temporal reasoning from tables in a domain that does not contain temporal quantities.

3.9 Experiments and Results

The goal of our experiments is to study how well different modeling approaches address the INFOTABS data, and also to understand the impact of various artifacts on them. First, we will consider different approaches for representing tables in ways that are amenable to modern neural models.

3.9.1 Representing Tables

A key aspect of the INFOTABS task that does not apply to the standard NLI task concerns how premise tables are represented. As baselines for future work, let us consider several different approaches.

1. **Premise as Paragraph (Para)**: We convert the premise table into paragraphs using fixed template applied to each row. For a table titled `t`, a row with key `k` and value `v` is written as the sentence *The* `k` *of* `t` *are* `v`. For example, for the table in Figure 3.1, the row with key *Equipment* gets mapped to the sentence *The equipment of Dressage are horse, horse tack.* We have a small number of exceptions: e.g., if the key is *born* or *died*, we use the following template: `t` *was* `k` *on* `v`.

   The sentences from all the rows in the table are concatenated to form the premise paragraph. While this approach does not result in grammatical sentences, it fits the interface for standard sentence encoders.

2. **Premise as Sentence (Sent):** Since hypotheses are typically short, they may be derived from a small subset of rows. Based on this intuition, we use the word mover distance [133] to select the closest and the three closest sentences to the hypothesis from the paragraph representation (denoted by WMD-1 and WMD-3, respectively).
3. **Premise as Structure 1 (TabFact)**: Following [26], we represent tables by a sequence of `key : value` tokens. Rows are separated by a semi-colon and multiple values for the same key are separated by a comma.
4. **Premise as Structure 2 (TabAttn)**: To study an attention based approach, such as that of [193], we convert keys and values into a contextually enriched vectors by first converting them into sentences using the Para approach above, and applying a contextual encoder to each sentence. From the token embeddings, we obtain the embeddings corresponding of the keys and values by mean pooling over only those tokens.

3.9.2 Modeling Table Inferences

Based on the various representations of tables described above, we developed a collection of models for the table inference problem, all based on standard approaches for NLI. Due to space constraints, we give a brief description of the models here and refer the interested reader to the code repository for implementation details.

For experiments where premises are represented as sentences or paragraphs, we evaluated a feature-based baseline using unigrams and bigrams of tokens. For this model (referred to as *SVM*), we used the LibLinear library [62].

For these representations, we also evaluated a collection of BERT-class of models. Following the standard setup, we encoded the premise-hypothesis pair, and used the classification token to train a classifier, specifically a two-layer feedforward network that predicts the label. The hidden layer had half the size of the token embeddings. We compared RoBERTa$_L$ (Large), RoBERTa$_B$ (Base) and BERT$_B$ (Base) in our experiments.

We used the above BERT strategy for the TabFact representations as well. For the TabAttn representations, we implemented the popular decomposable attention model [193] using the premise key-value embeddings and hypothesis token embeddings with 512 dimensional attend and compare layers.

We implemented all our models using the PyTorch with the transformers library [281]. We trained our models using Adagrad with a learning rate of 10^{-4}, chosen by preliminary experiments, and using a dropout value of 0.2. All our results in the following sections are averages of models trained from three different random seeds.

3.9.3 Experimental Results

Our experiments answer a series of questions.

3.9.3.1 Does our dataset exhibit hypothesis bias?

Before we consider the question of whether we can model premise-hypothesis relationships, let us first see if a model can learn to predict the entailment label without using the premise, thereby exhibiting an undesirable artifact. We consider three classes of models to study hypothesis bias in INFOTABS.

Hypothesis Only (hypo-only): The simplest way to check for hypothesis bias is to train a classifier using only the hypotheses. Without a premise, a classifier should fail to correlate the hypothesis and the label. We represent the hypothesis in two ways a) using unigrams and bigrams for an SVM, and b) using a single-sentence BERT-class model. The results of the experiments are given in Table 3.9.

Dummy or Swapped Premise: Another approach to evaluate hypothesis bias is to provide an unrelated premise and train a full entailment model. We evaluated two cases, where every premise is changed to a (a) *dummy* statement (*to be or not to be*), or (b) a randomly *swapped* table that is represented as paragraph. In both cases, we trained a RoBERTa$_L$ classifier as described in §3.9.2. The results for these experiments are presented in Table 3.10.

Results and Analysis: Looking at the Dev and α_1 columns of Tables 3.9 and 3.10, we see that these splits do have hypothesis bias. All the BERT-class models discover such artifacts equally well. However, we also observe that the performance on α_2 and α_3 data splits is worse since the artifacts in the training data do not occur in these splits. We see a performance gap of $\sim$ 12% as compared to Dev and α_1 splits in all cases. While there is some hypothesis bias in these splits, it is much less pronounced.

An important conclusion from these results is that the baseline for all future models trained on these splits should be the best premise-free performance. From the results here,

these correspond to the *swapped* setting.

3.9.3.2 How do trained NLI systems perform on our dataset?

Given the high leaderboard accuracies of trained NLI systems, the question of whether these models can infer entailment labels using a linearization of the tables arises. To study this, we trained RoBERTa$_L$ models on the SNLI and MultiNLI datasets. The SNLI model achieves an accuracy of 92.56% on SNLI test set. The MultiNLI model achieves an accuracy of 89.0% on matched and 88.99% on the mismatched MultiNLI test set. We evaluate these models on the WMD-1 and the Para representations of premises.

Results and Analysis: In Table 3.11, all the results point to the fact that pre-trained NLI systems do not perform well when tested on INFOTABS. We observe that full premises slightly improve performance over the WMD-1 ones. This might be due to a) ineffectiveness of WMD to identify the correct premise sentence, and b) multi-row reasoning.

3.9.3.3 Does training on the paragraph/sentence representation of a premise help?

The next set of experiments compares BERT-class models and SVM trained using the paragraph (Para) and sentence (WMD-n) representations. The results for these experiments are presented in Table 3.12.

Results and Analysis: We find that training with the INFOTABS training set improves model performance significantly over the previous baselines, except for the simple SVM model which relies on unigrams and bigrams. We see that RoBERTa$_L$ outperforms its base variant and BERT$_B$ by around $\sim$ 9% and $\sim$ 14% respectively. Similar to the earlier observation, providing full premise is better than selecting a subset of sentences.

Importantly, α_2 and α_3 performance is worse than α_1, not only suggesting the difficulty of these data splits, but also showing that models overfit both lexical patterns (based on α_2) or domain-specific patterns (based on α_3).

3.9.3.4 Does training on premise encoded as structure help?

Rather than linearizing the tables as sentences, we can try to encode the structure of the tables. We consider two representative approaches for this, TabFact and TabAttn, each associated with a different model as described in §3.9.2. The results for these experiments

are listed in Table 3.13.

Results and Analysis: The idea of using this family of models was to leverage the structural aspects of our data. We find that the TabAttn model, however, does not improve the performance. We assume that this might be due to the bag of words style of representation that the classifier e mploys. We find, however, that providing premise structure information helps the TabFact model perform better than the RoBERTa$_L$+Para model. As before model performance drops for α_2 and α_3.

3.9.3.5 How many types of reasoning does a trained system predict correctly?

Using a RoBERTa$_L$, which was trained on the paragraph (Para) representation, we analyzed the examples in Dev and α_3 data splits that were annotated by experts for their types of reasoning (§3.8). Figures 3.8 and 3.9 show the summary of this analysis.

Results and Analysis: Figures 3.8 and 3.9 show the histogram of reasoning types among correctly predicted examples. Compared to Figures 3.4 and 3.5, we see a decrease in correct predictions across all reasoning types for both Dev and α_3 sets. In particular, in the Dev set, the model performs poorly for the knowledge & common sense, multi-row, coreference, and temporal reasoning categories.

3.9.3.6 Labelwise F1 score analysis

The F1 scores per label for two model baselines are in Table 3.14. We observe that neutral is easier than entailment and contradiction for both baseline, which is expected as neutrals are mostly associated with subjective/out-of-table reasonings which makes them syntactically different and easier to predict correctly. Despite this, we found that in all evaluations in (§7.6) (except for α_2 test set), our models found neutrals almost as hard as the other two labels, with only an $\sim$ 3% gap between the F-scores of the neutral label and the next best label. For α_2 test set neutral are much easier than entailment and contradiction. This is expected as entailment and contradiction in α_2 were adversarially flipped; hence, these predictions become remarkably harder compared to neutrals. Furthermore, α_3 is the hardest data split, followed by α_2 and α_1.

3.9.3.7 Discussion

Our results show that: 1) INFOTABS contains a certain amount of artifacts which transformer-based models learn, but all models have a large gap to human performance; and 2) models accuracies drop on α_2 and α_3, suggesting that all three results together should be used to characterize the model, and not any single one of them. All our models are significantly worse than the human performance (84.04%, 83.88% and 79.33% for α_1, α_2 and α_3 respectively). With a difference of $\sim$ 14% between our best model and the human performance, these results indicate that INFOTABS is a challenging dataset.

3.10 Conclusion

We presented a new high quality natural language inference dataset, INFOTABS, with heterogeneous semi-structured premises and natural language hypotheses. Our analysis showed that our data encompasses several different kinds of inferences. INFOTABS has multiple test sets that are designed to pose difficulties to models that only learn superficial correlations between inputs and the labels, rather than reasoning about the information. Via extensive experiments, we showed that derivatives of several popular classes of models find this new inference task challenging. We expect that the dataset can serve as a testbed for developing new kinds of models and representations that can handle semi-structured information as first class citizens.

Dressage	
Highest governing body	International Federation for Equestrian Sports (FEI)
	Characteristics
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, horse tack
Venue	Arena, indoor or outdoor
	Presence
Country or region	Worldwide
Olympic	1912
Paralympic	1996

H1: Dressage was introduced in the Olympic games in 1912.

H2: Both men and women compete in the equestrian sport of Dressage.

H3: A dressage athlete can participate in both individual and team events.

H4: FEI governs dressage only in the U.S.

Figure 3.1: A semi-structured premise (the table). Two hypotheses (H1, H2) are entailed by it, H3 is neither entailed nor contradictory, and H4 is a contradiction.

Kamloops	
Type	Elected city council
Mayor	Ken Christian
Governing body	Kamloops City Council
MP	Cathy McLeod
MLAs	Peter Milobar, Todd Stone

H1: Kamloops has a democracy structure.

H2: If Ken Christian resigns as Mayor of Kamloops then Cathy McLeod will most likely replace him.

H3: Kamloops is ruled by a president.

Jefferson Starship	
Origin	San Francisco California
Genres	Rock, hard rock, psychedelic rock, progressive rock, soft rock
Years active	1970 - 1984, 1992 - present
Labels	RCA Grunt Epic
Associated acts	Jefferson Airplane Starship, KBC Band, Hot Tuna
Website	www.jeffersonstarship.net

H1: Jefferson Starship was started on the West Coast of the United States.

H2: Jefferson Starship won many awards for its music.

H3: Jefferson Starship has performed continuously since the 1970s.

Figure 3.2: Two semi-structured premises (the tables), and three hypotheses (H1: entailment, H2: Neutral, and H3: contradiction) that correspond to each table.

Femme aux Bras Croisés	
Artist	Pablo Picasso
Year	1901-02
Medium	Oil on canvas
Dimensions	81 cm 58 cm (32 in 23 in)
Location	Privately held

Figure 3.3: An example premise.

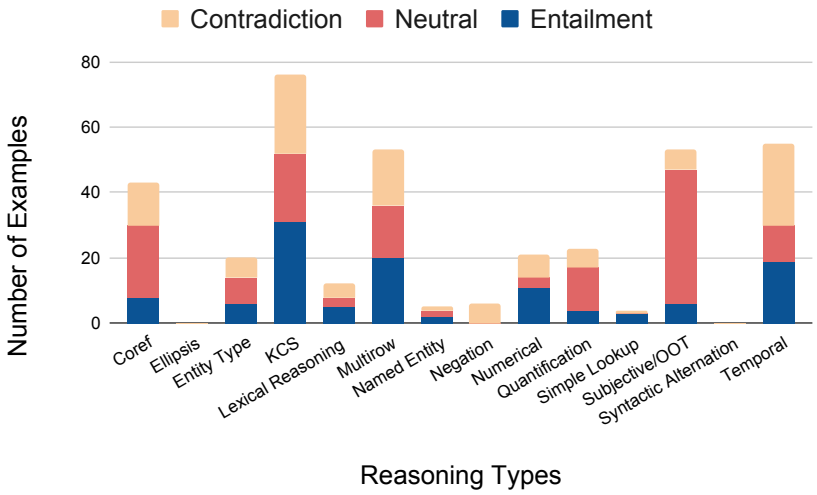


Figure 3.4: Number of examples per reasoning type in the dev set.

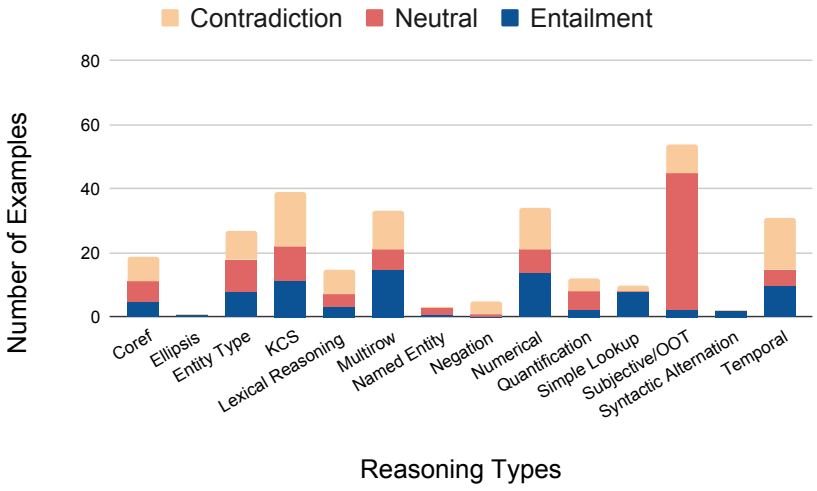


Figure 3.5: Number of examples per reasoning type in the α_3 set.

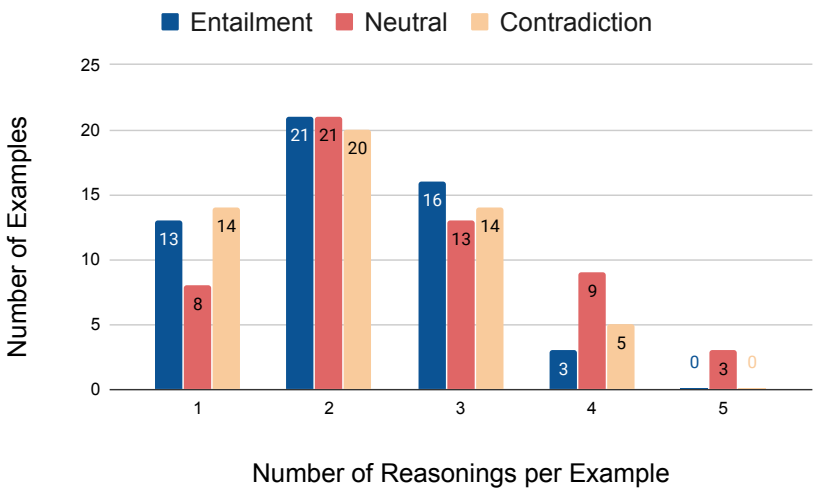


Figure 3.6: Number of reasonings per example in the dev set.

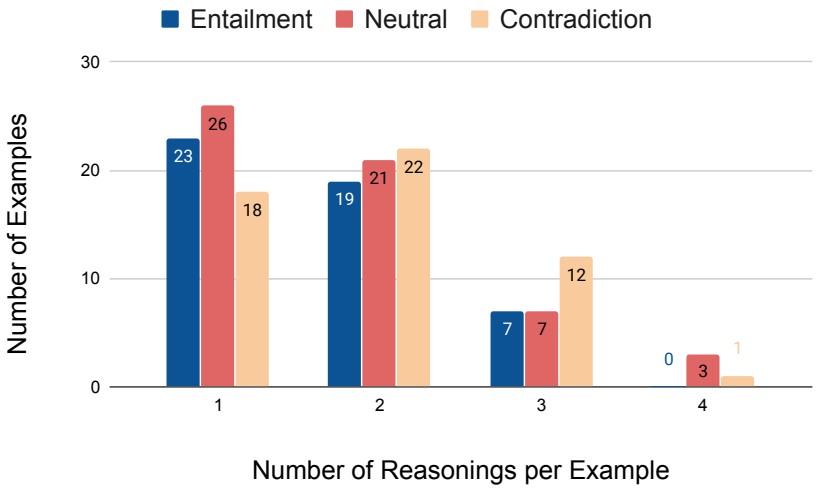


Figure 3.7: Number of reasonings per example in the α_3 set.

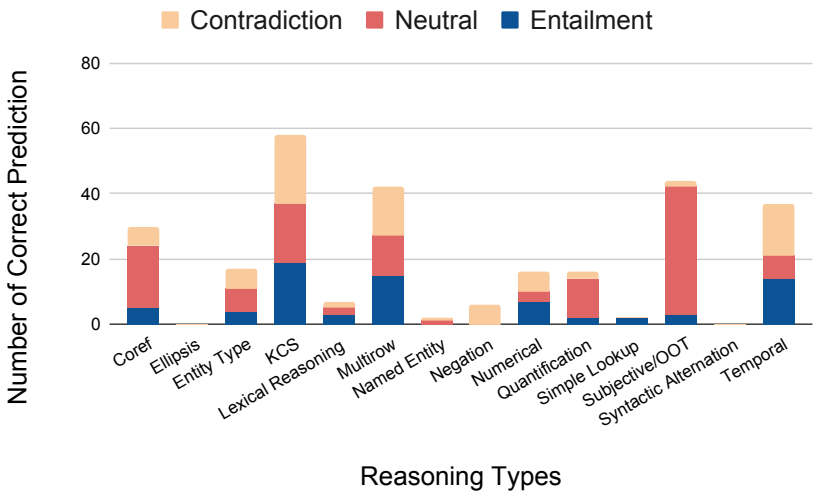


Figure 3.8: Number of correct predictions per reasoning type in the dev set.

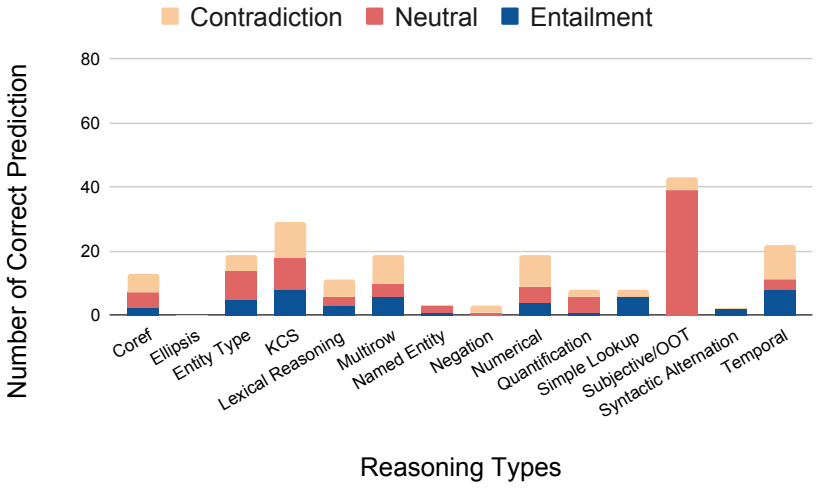


Figure 3.9: Number of correct predictions per reasoning type in the α_3 test set.

Table 3.1: Number of tables and premise-hypothesis pairs for each data split.

Data split	# tables	# pairs
Train	1740	16538
Dev	200	1800
α_1 test	200	1800
α_2 test	200	1800
α_3 test	200	1800

Table 3.2: Statistics of the premises and annotators across all discussed train-test splits.

Split	Train	Dev	α_1	α_2	α_3
Number of Unique Keys	1558	411	466	332	409
Number of Unique Keys Intersection with Train	-	334	312	273	94
Average # of keys per table	8.8	8.7	8.8	8.8	13.1
Number of Distinct Annotators	121	35	37	31	23
Annotator Intersection with Train	-	33	37	30	19
Number of Instances annotated by a Train annotator	-	1794	1800	1797	1647

Table 3.3: Mean length of the generated hypothesis sentences across all discussed train-test splits (standard deviation is in range 2.8 to 3.5).

Label	Train	Dev	α_1	α_2	α_3
Entail	9.80	9.71	9.90	9.33	10.5
Neutral	9.84	9.89	10.0	9.59	9.84
Contradict	9.37	9.72	9.84	9.40	9.86

Table 3.4: Mean statistic of the hypothesis sentences word overlapped with premises tables across all discussed train-test splits (standard deviation is in range 0.17 to 0.22).

Label	Train	Dev	α_1	α_2	α_3
Entail	0.52	0.47	0.45	0.46	0.48
Neutral	0.46	0.44	0.44	0.49	0.46
Contradict	0.44	0.43	0.45	0.44	0.46

Table 3.5: Categories for all data splits (excluding α_3) in percentage (%). Others (< 3%) include categories such as University, Event, Aircraft, Product, Game, Architecture, Planet, Awards, Wineyard, Airport, Language, Element, Car.

Category	**Train**	**Dev**	α_1	α_2	Category	Train	Dev	α_1	α_2
Person	23.68	27	28.5	35.5	Musician	14.66	19	18.5	22.5
Movie	10.17	10	9	11.5	Album	9.08	7	3.5	4.5
City	8.05	8.5	8	7	Painting	5.98	4.5	4	3.5
Organization	4.14	2	1	0.5	Food / Drinks	4.08	4	4	3
Country	3.74	6	9	3.5	Animal	3.56	4.5	4	4
Sports	4.6	3.5	2.5	0.0	Book	2.18	0.5	3	2.5
Other	6.07	8.00	5.00	2.00					

Table 3.6: Categories for α_3 datasplit. Others (< 3%) include categories such as Computer, Occupation, Restaurant, Engines, Equilibrium, OS, Cloud, Bus/Train Station, Coffee House, Cars, Bus/Train Provider, Hotel, Math, Flight.

Category	α_3 (%)	**Category**	α_3 (%)
Diseases	20.4	Festival	17.41
Bus / Train Lines	14.93	Exams	8.46
Element	4.98	Exams	8.46
Bridge	3.98	Disasters	3.48
Smartphone	3.48	Other	18.9

Table 3.7: Inter-annotator agreement statistics.

Dataset	**Cohen's Kappa**	**Human Accuracy**	**Majority Agreement**
Dev	0.78	79.78	93.52
α_1	0.80	84.04	97.48
α_2	0.80	83.88	96.77
α_3	0.74	79.33	95.58

Table 3.8: Exact, Individual and Kappa values for verification's statistics.

Exact agreement between annotators.		
Dataset	**Number**	**Gold/Total**
Dev	3	350 / 469
	4	529 / 601
	5	550 / 605
	no agreement	116
α_1	3	184 / 292
	4	459 / 533
	5	863 / 922
	no agreement	45
α_2	3	245 / 348
	4	453 / 537
	5	812 / 857
	no agreement	58
α_2	3	273 / 422
	4	441 / 524
	5	706 / 765
	no agreement	79

Individual agreement with gold / majority label.		
Dataset	**Statistics**	**Agreement (%)**
Dev	Gold	71.12
	Majority	81.65
α_1	Gold	78.52
	Majority	87.24
α_2	Gold	77.74
	Majority	86.32
α_3	Gold	73.22
	Majority	84.01
Average	Gold	75.15
	Majority	84.8

Kappa values across splits		
Dataset	**Fleiss**	**Cohen**
Dev	0.4601	0.7793
α_1	0.6375	0.7930
α_2	0.5962	0.8001
α_3	0.5421	0.7444

Table 3.9: Accuracy of hypothesis-only baselines on the INFOTABS Dev and test sets.

Model	Dev	α_1	α_2	α_3
Majority	33.33	33.33	33.33	33.33
SVM	59.00	60.61	45.89	45.89
BERT$_B$	62.69	63.45	49.65	50.45
RoBERTa$_B$	62.37	62.76	50.65	50.8
RoBERTa$_L$	60.51	60.48	48.26	48.89

Table 3.10: Accuracy with dummy/swapped premises.

Premise	Dev	α_1	α_2	α_3
dummy	60.02	59.78	48.91	46.37
swapped	62.94	65.11	52.55	50.21

Table 3.11: Accuracy of test splits with structured representation of premises with RoBERTa$_L$ trained on SNLI and MultiNLI training data.

Premise	Dev	α_1	α_2	α_3
Trained on SNLI				
WMD-1	49.44	47.5	49.44	46.44
Para	54.44	53.55	53.66	46.01
Trained on MultiNLI				
WMD-1	44.44	44.67	46.88	44.01
Para	55.77	53.83	55.33	47.28

Table 3.12: Accuracy of paragraph and sentence premise representation reported on SVM, BERT$_B$, RoBERTa$_B$ and RoBERTa$_L$.

Premise	Dev	α_1	α_2	α_3
Train with SVM				
Para	59.11	59.17	46.44	41.28
Train with BERT$_B$				
Para	63.00	63.54	52.57	48.17
Train with RoBERTa$_B$				
Para	67.2	66.98	56.87	55.36
Train with RoBERTa$_L$				
WMD-1	65.44	65.27	57.11	52.55
WMD-3	72.55	70.38	62.55	61.33
Para	75.55	74.88	65.55	64.94

Table 3.13: Accuracy on structured premise representation reported on BERT_B, RoBERTa_B and RoBERTa_L.

Premise	Dev	α_1	α_2	α_3
Train with BERT_B				
TabFact	63.67	64.04	53.59	49.05
Train with RoBERT_B				
TabFact	68.06	66.7	56.87	55.26
Train with RoBERTa_L				
TabAttn	63.63	62.94	49.37	49.04
TabFact	77.61	75.06	69.02	64.61

Table 3.14: F1 Score (%) with various baselines. All models are trained with RoBERTa_L.

Premise as Paragraph			
Split	Entailment	Neutral	Contradiction
Dev	76.19	79.02	72.73
α_1	74.69	77.85	69.85
α_2	57.06	80.36	62.14
α_3	65.27	66.06	61.61
Premise as TabFact			
Split	Entailment	Neutral	Contradiction
Dev	77.69	79.45	74.77
α_1	76.43	80.34	73.07
α_2	55.34	80.83	64.44
α_3	65.92	67.28	63.57

CHAPTER 4

KNOWLEDGE INTEGRATION PRE-PROCESSING

Adapted from J. Neeraja, V. Gupta, V. Srikumar, *Incorporating external knowledge to enhance tabular reasoning*, in Proceedings of the 2021 Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 6–11, 2021, Association for Computational Linguistics, pp. 2799–2809.

In this Chapter, we argue that instead of relying on the neural network to *"magically"* work for tabular structures, as done in Chapter 3, we should carefully think about the representation of semi-structured data, and the incorporation of both implicit and explicit knowledge into neural models. We highlights that simple pre-processing steps are important, especially for better generalization. Using the InfoTabS dataset as discussed in Chapter 3, we present a focused study that investigates (a) the poor performance of existing models, (b) connections to information deficiency in the tabular premises, and, (c) simple yet effective mitigations for these problems.

We use the table and hypotheses in Figure 4.1 as a running example through this chapter, and refer to the left column as its keys.[1] Tabular inference is challenging for several reasons: (a) **Poor table representation**: The table does not explicitly state the relationship between the keys and values. (b) **Missing implicit lexical knowledge** due to limited training data: This affects interpreting words like *'fewer'*, and *'over'* in H1 and H2 respectively. (c) **Presence of distracting information**: All keys except *No. of listings* are unrelated to the hypotheses H1 and H2. (d) **Missing domain knowledge about keys**: We need to interpret the key *Volume* in the financial context for this table.

In the absence of large labeled corpora, any modeling strategy needs to explicitly address these problems. In this chapter, we propose effective approaches for addressing

[1]Keys in the InfoTabS tables are similar to column headers in the TabFact database-style tables.

them, and show that they lead to substantial improvements in prediction quality, especially on adversarial test sets. We recommend that these pre-processing steps should be standardized across table reasoning tasks.[2]

4.1 Contributions

This chapter makes the following contributions:

1. We analyse why the existing state-of-the-art BERT class models struggle on the challenging task of NLI over tabular data.
2. We propose solutions to overcome these challenges via simple modifications to inputs using existing language resources.
3. Through extensive experiments, we show significant improvements to model performance, especially on challenging adversarial test sets.

This work is published at NAACL 2021 as [182].

4.2 Background

There have been many works which study several NLP tasks on semi-structured tabular data. These include tabular NLI and fact verification tasks such as TabFact [26], and InfoTabS [84], various question answering and semantic parsing tasks [1, 27, 128, 151, 195, 245], and table-to-text generation and its evaluation [194, 207]. Several, models for better representation of tables such as TAPAS [94], TaBERT [291], and TabStruc [300] were recently proposed. [294, 295] and [58] study pre-training for improving tabular inference, similar to our MutliNLI pre-training. However, much of these recent work focuses on building sophisticated neural models, without explicit focus on how these models (designed for raw text) adapt to the tabular data. The proposed modifications we proposed in this chapter are simple and intuitive. Yet, existing table reasoning papers have not studied the impact of such input modifications.

4.3 Challenges and Proposed Solutions

We examine the issues highlighted above and propose simple solutions to mitigate them below.

[2]The updated dataset, along with associated scripts, are available at https://knowledge-infotabs.github.io.

4.3.1 Better Table Representation (BTR)

One way to represent the premise table is to use a universal template to convert each row of the table into sentence which serves as input to a BERT-style model. [84] suggest that in a table titled `t`, a row with key `k` and value `v` should be converted to a sentence using the template: "The `k` of `t` are `v`." Despite the advantage of simplicity, the approach produces ungrammatical sentences. In our example, the template converts the *Founded* row to the sentence *"The Founded of New York Stock Exchange are May 17, 1792; 226 years ago."*.

We note that keys are associated with values of specific entity types such as MONEY, DATE, CARDINAL, and BOOL, and the entire table itself has a category. Therefore, we propose type-specific templates, instead of using the universal one.[3] In our example, the table category is *Organization* and the key *Founded* has the type DATE. A better template for this key is "`t` was `k` on `v`", which produces the more grammatical sentence *"New York Stock Exchange was Founded on May 17, 1792; 226 years ago."*. Furthermore, we observe that including the table category information i.e. *"New York Stock Exchange is an Organization."* helps in better premise context understanding.[4]

4.3.2 BPR Templates

Here, we are listing down some of the diverse example templates we have framed.

- For the table category *Bus/Train Lines* and key *Disabled access* with BOOL value YES, follow template: "*t* has *k*."

Original Premise Sentence	*"The Disabled access of Tukwila International Boulevard Station are Yes."*
BPR Sentence	*"Tukwila International Boulevard Station has Disabled access."*

- For the table category *Movie* and key *Box office* with MONEY type, follow template: "In the *k*, *t* made *v*."

Original Premise Sentence	*"The Box office of Brokeback Mountain are $178.1 million."*
BPR Sentence	*"In the Box office, Brokeback Mountain made $178.1 million."*

- For the table category *City* and key *Total* with CARDINAL type, follow template: "The *k* area of *t* is *v*."

[3] The construction of the template sentences based on entity type is a one-time manual step.

[4] This category information is provided in the InfoTabS and TabFact datasets. For other datasets, it can be inferred easily by clustering over the keys of the training tables.

Original Premise Sentence *"The Total of Cusco are 435,114."*
BPR Sentence *"The Total area of Cusco is 435,114."*

- For the table category *Painting* and key *Also known as*, follow template: "The k area of t is v."

Original Premise Sentence *"The Also known as of Et in Arcadia ego are Les Bergers d'Arcadie."*
BPR Sentence *"Et in Arcadia ego is Also known as Les Bergers d'Arcadie."*

- For the table category *Person* and key *Died* with DATE type , follow template: "t k on v."

Original Premise Sentence *"The Died of Jesse Ramsden are November 1800 (1800-11-05) (aged 65) Brighton, Sussex."*
BPR Sentence *"Jesse Ramsden Died on 5 November 1800 (1800-11-05) (aged 65) Brighton, Sussex."*

4.3.3 Implicit Knowledge Addition (KG Implicit)

Tables represent information *implicitly*; they do not employ connectives to link their cells. As a result, a model trained only on tables struggles to make lexical inferences about the hypothesis, such as the difference between the meanings of *'before'* and *'after'*, and the function of negations. This is surprising, because the models have the benefit of being pre-trained on large textual corpora.

Recently, [6] and [205] showed that we can pre-train models on specific tasks to incorporate such implicit knowledge. [58] use pre-training on synthetic data to improve the performance on the TabFact dataset. Inspired by these, we first train our model on the large, diverse and *human-written* MultiNLI dataset. Then, we fine tune it to the InfoTabS task. Pre-training with MultiNLI data exposes the model to diverse lexical constructions. Furthermore, it increases the training data size by 433K (MultiNLI) example pairs. This makes the representation better tuned to the NLI task, thereby leading to better generalization.

4.3.4 Distracting Rows Removal (DRR)

Not all premise table rows are necessary to reason about a given hypothesis. In our example, for the hypotheses H1 and H2, the row corresponding to the key *No. of listings* is sufficient to decide the label for the hypothesis. The other rows are an irrelevant distraction. Further, as a practical concern, when longer tables are encoded into sentences as described above, the resulting number of tokens is more than the input size restrictions of

existing models, leading to useful rows potentially being cropped. Therefore, it becomes important to prune irrelevant rows.

To identify relevant rows, we employ a simplified version of the alignment algorithm used by [287, 288] for retrieval in reading comprehension.

First, every word in the hypothesis sentence is aligned with the most similar word in the table sentences using cosine similarity. We use fastText [115, 168] embeddings for this purpose, which preliminary experiments revealed to be better than other embeddings. Then, we rank rows by their similarity to the hypothesis, by aggregating similarity over content words in the hypothesis. [287] used inverse document frequency for weighting words, but we found that simple stop word pruning was sufficient. We took the top k rows by similarity as the pruned representative of the table for this hypothesis. The hyper-parameter k is selected by tuning on a development set.

4.3.4.1 fastText representation

For word representation, [287] have used BERT and Glove embeddings. In our case, we prefer to use fastText word embeddings over Glove because fastText embedding uses sub-word information which helps in capturing different variations of the context words. Furthermore, fastText embeddings is also as better choice than BERT for our task because 1. Firstly, we are embedding single sentential form of diverse rows instead of longer context similar paragraphs, 2. Secondly, all words (especially keys) of the rows across all the tables are used only in one context, whereas BERT is useful when same word is used with different contexts across paragraphs, 3. Thirdly, in all tables, the number sentences to select from is bounded by maximum rows in the table, which is a small number (8.8 in train, dev, α_1, α_2 and 13.1 in α_3), and 4. Lastly, using fastText is much faster to compute than BERT for obtaining embeddings.

4.3.4.2 Binary weighting scheme

Since, we are embedding single sentential form of diverse rowsinstead of longer context related paragraphs, we found that using binary weighting 0 for stop words and 1 for others is more effective than the idf weighting, which is useful only for longer paragraph context with several lexical terms.

4.3.5 Explicit Knowledge Addition (KG Explicit)

We found that adding *explicit* information to enrich keys improves a model's ability to disambiguate and understand them. We expand the pruned table premises with contextually relevant key information from existing resources such as WordNet (definitions) or Wikipedia (first sentence, usually a definition).[5]

To find the best expansion of a key, we use the sentential form of a row to obtain the BERT embedding (on-the-fly) for its key. We also obtain the BERT embeddings of the same key from WordNet examples (or Wikipedia sentences).[6] Finally, we concatenate the WordNet definition (or the Wikipedia sentence) corresponding to the highest key embedding similarity to the table. As we want the contextually relevant definition of the key, we use the BERT embeddings rather than non-contextual ones (e.g., fastText). For example, the key *volume* can have different meanings in various contexts. For our example, the contextually best definition is *"In capital markets, **volume**, is the total number of a security that was traded during a given period of time."* rather than the other definition *"In thermodynamics, the **volume** of a system is an extensive parameter for describing its thermodynamic state."*.

4.4 Experiment and Analysis

Our experiments are designed to study the research question: *Can today's large pre-trained models exploit the information sources described in §4.3 to better reason about tabular information?*

4.4.1 Experimental Setup

4.4.1.1 Datasets

Our experiments uses InfoTabS, a tabular inference dataset from [84]. The dataset is heterogeneous in the types of tables and keys, and relies on background knowledge and common sense. Unlike the TabFact dataset [26], it has all three inference labels, namely entailment, contradiction and neutral. Importantly, for the purpose of our evaluation, it has three test sets. In addition to the usual development set and the test set (called α_1), the

[5]Usually multi-word keys are absent in WordNet, in this case we use Wikipedia. The WordNet definition of each word in the key is used if the multi-word key is absent in Wikipedia.

[6]We prefer using WordNet examples over definition for BERT embedding because (a) an example captures the context in which key is used, and (b) the definition may not always contain the key tokens.

dataset has two adversarial test sets: a contrast set α_2 that is lexically similar to α_1, but with minimal changes in the hypotheses and flip entail-contradict label, and a zero-shot set α_3 which has long tables from different domains with little key overlap with the training set.

4.4.1.2 Models

For a fair comparison with earlier baselines, we use RoBERTa-large (RoBERTa$_L$) for all our experiments. We represent the premise table by converting each table row into a sentence, and then appending them into a paragraph, i.e. the *Para* representation of [84].

4.4.1.3 Hyperparameters settings

For the distracting row removal (+DRR) step, we have a hyper-parameter k. We experimented with $k \in \{2, 3, 4, 5, 6\}$, by predicting on +DRR development premise on model trained on Original training set (i.e. BTR), as shown in Table 4.1. The development accuracy increases significantly as k increases from 2 to 4 and then from 4 to 6, increases marginally (1.5% improvement). Since our goal is to remove distracting rows, we use the lowest hyperparameter with good performance i.e. $k = 4$.[7].

4.4.2 Results and Analysis

Table 4.2 shows the results of our experiments.

4.4.2.1 BTR

As shown in Table 4.2, with BTR, we observe that the RoBERTa$_L$ model improves performance on all dev and test sets except α_3. There are two main reasons behind this poor performance on α_3.

First, the zero-shot α_3 data includes unseen keys. The number of keys common to α_3 and the training set is 94, whereas for, dev, α_1 and α_2 it is 334, 312, and 273 respectively (i.e., 3-5 times more). Second, despite being represented by better sentences, due to the input size restriction of RoBERTa$_L$ some relevant rows are still ignored.

[7]Indeed, the original InfoTabS work points out that no more than four rows in a table are needed for any hypothesis.

4.4.2.2 KG implicit

We observe that *implicit* knowledge addition via MNLI pre-training helps the model reason and generalize better. From Table 4.2, we can see significant performance improvement in the dev and all three test sets.

4.4.2.3 DRR

This leads to significant improvement in the α_3 set. We attribute this to two primary reasons: First, α_3 tables are longer (13.1 keys per table on average, versus 8.8 keys on average in the others), and DRR is important to avoid automatically removing keys from the bottom of a table due to the limitations in RoBERTa$_L$ model's input size. Without these relevant rows, the model incorrectly predicts the neutral label. Second, α_3 is a zero-shot dataset and has significant proportion of unseen keys which could end up being noise for the model. The slight decrease in performance on the dev, α_1 and α_2 sets can be attributed to model utilising spurious patterns over irrelevant keys for prediction.[8] We validated this experimentally by testing the original premise trained model on the DRR test tables. Table 4.3 shows that without pruning, the model focuses on irrelevant rows for prediction.

4.4.2.4 KG explicit

With *explicit* contextualized knowledge about the table keys, we observe a marginal improvement in dev, α_1 test sets and a significant performance gain on the α_2 and α_3 test sets. Improvement in the α_3 set shows that adding external knowledge helps in the zero-shot setting. With α_2, the model can not utilize spurious lexical correlations[9] due to its adversarial nature, and is forced to use the relevant keys in the premise tables, thus adding explicit information about the key improves performance more for α_2 than α_1 or dev. Appendix A shows some qualitative examples.

---

[8]Performance drop of dev and α_2 is also marginal i.e. (dev: 79.57 to 78.77, α_1: 78.27 to 78.13, α_2: 71.87 to 70.90), as compared to InfoTabS WMD-top3 i.e (dev: 75.5 to 72.55,α_1: 74.88 to 70.38, α_2: 65.44 to 62.55), here WMD-top3 performance numbers are taken from [84].

[9]The hypothesis-only baseline for α_2 is 48.5% versus α_1: 60.5 % and dev: 60.5 % [84].

4.4.3 Ablation Study

We perform an ablation study as shown in Table 4.4, where instead of doing all modification sequentially one after another (+), we do only one modification at a time to analyze its effects.

Through our ablation study we observe that: (a) **DRR** improves performance on the dev, α_1, and α_2 sets, but slightly degrades it on the α_3 set. The drop in performance on α_3 is due to spurious artifact deletion as explained in details in later Section 4.4.3.2. (b) **KG explicit** gives performance improvement in all sets. Furthermore, there is significant boost in performance of the adversarial α_2 and α_3 sets.[10] (c) Similarly, **KG implicit** shows significant improvement in all test sets. The large improvements on the adversarial sets α_2 and α_3 sets, suggest that the model can now reason better. Although, implicit knowledge provides most performance gain, all modifications are needed to obtain the best performance for all sets (especially on the α_3 set). We show in Section 4.4.3.1, Table 4.5, that implicit knowledge addition to a non-sentential table representation i.e. Struc [26, 84] leads to performance improvement as well.

4.4.3.1 TabFact representation experiment

Table 4.5 is implicit knowledge addition effect on non-para *Struc* representation i.e. a key value linearize representation as "key `k` : value `v`", rows separated by semicolon ";" [26, 84]. Here too the implicit knowledge addition leads to improvement in performance on all the sets.

4.4.3.2 Artifacts and model predictions

In Table 4.6 we show percentage of example which were corrected after modification and vice versa. Surprisingly, there is a small percentage of examples which are predicted correctly earlier with original premise (Para) but predicted wrongly after all the modifications (Mod), although such examples are much lesser than opposite case. We suspect that earlier model was also relying on spurious pattern (artifacts) for correct prediction on these examples earlier, which are now corrupted after the proposed modifications. Hence, the new model struggle to predict correctly on such examples.

[10] The KG explicit step is performed only for relevant keys (after DRR).

4.4.3.3 Hyperparameters k versus test-sets accuracy

We also trained a model both train and tested on the DRR table premise for increasing values of the hyper parameter k, as shown in Table 4.7. We also test the model trained on the entire para on pruned para with increasing value of hyperparameters $k \in \{2, 3, 4, 5, 6\}$ for the test sets α_1, α_2, and α_3. In all cases, except α_3, the performance with larger k is better. The increase in performance, even with $k > 4$, shows that the model is using more then required keys for prediction. Thus, the model is utlising the spurious pattern in irrelevant rows for the prediction.

In Appendix A, we also shows qualitative examples, where modification helps model predict correctly. We also provide some examples via distracting row removal modification, where model fails after modification.

4.5 Conclusion and Future Work

We introduced simple and effective modifications that rely on introducing additional knowledge to improve tabular NLI. These modifications governs what information is provided to a tabular NLI and how the given information is presented to the model. We presented a case study with the recently published InfoTabS dataset and showed that our proposed changes lead to significant improvements. Furthermore, we also carefully studied the effect of these modifications on the multiple test-sets, and why a certain modification seems to help a particular adversarial set.

We believe that our study and proposed solutions will be valuable to researchers working on question answering and generation problems involving both tabular and textual inputs, such as tabular/hybrid question answering and table-to-text generation, especially with difficult or adversarial evaluation. Looking ahead, our work can be extended to include explicit knowledge for hypothesis tokens as well. To increase robustness, we can also integrate structural constraints via data augmentation through NLI training. Moreover, we expect that structural information such as position encoding could also help better represent tables.

New York Stock Exchange	
Type	Stock exchange
Location	New York City, New York, U.S.
Founded	May 17, 1792; 226 years ago
Currency	United States dollar
No. of listings	2,400
Volume	US$20.161 trillion (2011)

H1: NYSE has fewer than 3,000 stocks listed.

H2: Over 2,500 stocks are listed in the NYSE.

H3: S&P 500 stock trading volume is over $10 trillion.

Figure 4.1: A tabular premise example. The hypotheses H1 is entailed by it, H2 is a contradiction and H3 is neutral i.e. neither entailed nor contradictory.

Table 4.1: Dev accuracy on increasing hyperparameter k.

Train	Dev	k = 2	k = 3	k = 4	k = 5	k = 6
BTR	DRR	71.72	74.83	77.50	78.50	79.00

Table 4.2: Accuracy with the proposed modifications on the Dev and test sets. Here, + represents the change with respect to the previous row. Reported numbers are the average over three random seed runs with standard deviation of 0.33 (+KG explicit), 0.46 (+DRR), 0.61 (+KG implicit), 0.86 (BTR), over all sets. All improvements are statistically significant with $p < 0.05$, except α_1 for BTR representation w.r.t to Para (Original). Here the Human and Para results are taken from Chapter 3.

Premise	Dev	α_1	α_2	α_3
Human	**79.78**	**84.04**	**83.88**	**79.33**
Para	75.55	74.88	65.55	64.94
BTR	76.42	75.29	66.50	64.26
+KG implicit	**79.57**	78.27	71.87	66.77
+DRR	78.77	78.13	70.90	68.98
+KG explicit	79.44	**78.42**	**71.97**	**70.03**

Table 4.3: Accuracy of model trained with original table but tested with DRR table with increasing hyper parameter k on all test sets.

k	α_1	α_2	α_3
2	71.44	67.33	64.83
3	75.05	69.33	67.33
4	77.72	69.83	68.22
5	77.77	70.28	**69.28**
6	**77.77**	**70.77**	69.22

Table 4.4: Ablation results with individual modifications.

Premise	**Dev**	α_1	α_2	α_3
Para	75.55	74.88	65.55	64.94
DRR	76.39	75.78	67.22	64.88
KG explicit	77.16	75.38	67.88	65.50
KG implicit	**79.06**	**78.44**	**71.66**	**67.55**

Table 4.5: Accuracy on InfoTabS data for Struc representation of tables. Here, + represents the change with respect to the previous row.

Premise	**Dev**	α_1	α_2	α_3
Struc	77.61	75.06	69.02	64.61
+ KG implicit	**79.55**	**78.66**	**72.33**	**70.44**

Table 4.6: Correct versus Incorrect predictions for Para model [84] and the model after the modifcations (Mod).

Para	**Mod**	**Dev**	α_1	α_2	α_3
✓	×	6.77	7.83	9.27	10.01
×	✓	**10.94**	**12.55**	**14.33**	**16.05**

Table 4.7: Dev accuracy with increasing hyper parameter k trained with both BPR and +DRR table.

Train	**Dev**	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**
+DRR	+DRR	77.61	77.94	78.16	78.38	79.00
BPR	+DRR	71.72	74.83	77.50	78.50	79.00

CHAPTER 5

KNOWLEDGE INTEGRATION TRANS-KBLSTM

Adapted from Y. Varun, A. Sharma, V. Gupta, *Trans-KBLSTM: An external knowledge enhanced transformer BiLSTM model for tabular reasoning*, in Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Dublin, Ireland and Online, May 27, 2022, Association of Computational Linguistics, pp. 62-78.

Chapter 4 highlight the significance of adding world knowledge for the tabular inference task (c.f. Table 5.1). The approach develops a knowledge addition strategy, namely *KG Explicit*, which expands the keys of a tabular premise with its definitions obtained from Wordnet and Wikipedia articles. These definitions are appended as a suffix to the original input as additional context. With this added additional knowledge, the model outperforms the original baseline. Despite improved effectiveness, this way of knowledge addition has the following drawbacks: (a) **Knowledge Extraction.** *KG Explicit* disambiguates multiple key definitions using the table context, ignoring the hypothesis content entirely. Additionally, the extended definition contains hypothesis-unrelated and unnecessary additional functional terms. All of these factors contribute to erroneous key-sense disambiguation and additional noise. (b) **Knowledge Addition.** *KG Explicit* adds knowledge by appending a suffix definition to existing inputs instead of using more effective semantic representations such as Knowledge Embedding (Graph Embedding or Learned representations). (c) **Knowledge Integration.** Finally, utilizing tokenized input BERT [51] to fuse word-pair relations yields considerably weaker semantic linkages between premise, hypothesis, and the external knowledge.

In this chapter, we propose a solution to above issues. We drew inspiration from [23] and utilize relational connections between premise and hypothesis to extract important

knowledge relations from ConceptNet [243] and Wordnet [170]. This enhancement reduces noise in knowledge addition, resulting in improved **Knowledge Extraction**. We embed relational terms in sentences using sentence transformers [217] to encode semantic representations of the relation, comparable to [68], culminating in successful **Knowledge Addition**. Finally, for effective **Knowledge Integration**, we combine these relational embeddings into a word-level language model, using BiLSTM [96], and backpropagate using our proposed BiLSTM and transformer architecture together to enhance model inferencing capabilities.

Our proposed model, Trans-KBLSTM, outperforms the earlier baseline, i.e., *KG Explicit* in full as well as limited supervision setting, substantially for some specific categories. Furthermore, knowledge addition via Trans-KBLSTM improve model *lexical*, *multi-row* and *Numerical* reasoning. We also performed a detailed ablation study to understand the importance of each component. This work is published at DeeLIO 2022 workshop at ACL 2022 as [262]. [1]

5.1 Contributions

The main contributions we make here are:

1. We address the challenges inherent in existing techniques, e.g., KG Explicit, for explicit knowledge addition in tabular reasoning.
2. We investigate a more efficient knowledge extraction method that involves using knowledge embeddings rather than directly appending them to the input.
3. We propose a novel architecture, namely Trans-KBLSTM, for integrating word-level knowledge effectively with BiLSTM's encoders with state-of-the-art transformers such as BERT.
4. Through extensive experiments, analysis and ablation studies, we demonstrate that Trans-KBLSTM improves reasoning for INFOTABS dataset.

5.2 Background: Knowledge Integration

Traditional approaches to integrating external knowledge into deep learning models do not use contextual embeddings from pre-trained language models. The Knowledge-

[1]The dataset, and associated scripts, are available at https://trans-kblstm.github.io/.

based Inference Model (KIM) [23] incorporates lexical relations (such as antonyms and synonyms) into the premise and hypothesis representations using attention and composition units. [150] provides a method to mine and exploit commonsense knowledge by defining inference rules between elements under different kinds of commonsense relations, with an inference cost for each rule. KG-Augmented Entailment System (KES) [118] augments the NLI model with external knowledge encoded using graph convolutional networks. ConseqNet [274] concatenates the output of the text-based model and the graph-based model and then feeds it to a classifier. [149] uses LSTMs and a novel knowledge-aware graph network module named KagNet to achieve state-of-the-art performance on CommonSenseQA. BiCAM [69] models incorporate knowledge from ConceptNet and AristoTuple KGs [46] by factorized bilinear pooling to improve performance on NLI Datasets.

Incorporating external knowledge into language models has been extensively explored in recent times. Approaches similar to the Tok-KTrans baseline described in §5.4.1 where external knowledge is added at input level were explored in [28, 174, 285]. At the representational level, the model understands these external knowledge additions and interacts with these representations using multi-head attention modules [21]. Other approaches include, pretraining on external knowledge corpus to inject knowledge [199, 260, 273], better knowledge representations [12], modifications to multi-head attention in pre-trained language models [89, 140], designing relation-aware tasks [283] and integration of knowledge through multi-head attention [68].

Recently, [143] finds that when explicit knowledge is added in the form of word-pair information, models such as [23] improve performance. However, such models necessitate the use of classic *seq2seq* architectures such as BiLSTM to integrate word-level knowledge. The use of external knowledge into Tabular data was first explored by us in Chapter 4 through *KG-Explicit* model described. We aim to improve on this benchmark through this extensive study. In our proposed approach, external knowledge is separately added to the premise and hypothesis using a multi-head attention dot product. To encode the contextual relationships between premise and hypothesis, we use a pre-trained language model, RoBERTa [158]. We combine the LM embeddings [68] and BiLSTM embeddings using a skip connection which preserves the premise-hypothesis relational context and

integrates knowledge effectively.

5.3 Proposed Trans-KBLSTM Model

We highlight the main model components and their implementation details in this section. We begin with a description of the knowledge relations retrieval technique, followed by a discussion of the model architecture's core components.

5.3.1 External Knowledge Relations Retrieval

It is challenging to retrieve contextually relevant knowledge relations from the knowledge graphs. The challenge is to retrieve task-relevant knowledge relations from massive volumes of noisy Knowledge Graph data. Our method is inspired by [23], which considers a connection to be significant if the knowledge graph contains the term pair relations.

5.3.1.1 Relational connections

We define relational connections between two sentences through external relational knowledge between each pair of words in the sentences. The token level relation connections are based on word triples derived from the knowledge graphs.

5.3.1.2 Relational connections retrieval

Stop words and punctuation are first removed from the premise and hypothesis. Then, we analyze the knowledge relational connections between the premise and hypothesis token pairs and compute the relationship attention matrix, A^r_{ij}, as follows:

$$A^r_{ij} = \begin{cases} 1 & i^{th} \text{ and } j^{th} \text{ words are } \textbf{related} \\ 0 & i^{th} \text{ and } j^{th} \text{ words are } \textbf{not related} \end{cases}$$

Each knowledge relational triple, consisting of two token terms (one from each premise and hypothesis) and their respective relationship is transformed into a complete grammatical sentence. For instance, the triple {Day, *Antonym*, Night} is transformed into "*Day* is the *opposite* of *Night*". For a complete list of knowledge templates refer to Appendix B. We utilize sentence transformers, as presented in [217], to convert the relationship phrase e.g. *"is opposite of"* in the preceding example into high-level semantic representations. The contextual representations denote the relational pair's across relational pairs.

5.3.1.3 Relational connection embedding

The contextual knowledge connections between premise and hypothesis token pairs are used to generate a relational vector, R_{ijk}. Each marginal vector R_{ij} is the k dimension BERT representation for the *"Relation Connection Sentence"* in the previously described sentential form constructed using the relationship between the i^{th} premise word and the j^{th} hypothesis word. For words whose relations are absent from knowlege source, we initialize the R_{ij} vector with 'zero' values.[2]

5.3.2 Model Architecture Details

Next, we described several components of our proposed model. Figure 5.1 describes the high level architecture of the **Trans-KBLSTM** model.

5.3.2.1 Transformer

We encode the premise and hypothesis using RoBERTa[158] to generate contextual word embeddings. Consider $P = \{p_i\}_{i=1}^{m}$ as table premise of length m and $H = \{h_j\}_{j=1}^{n}$ as hypothesis of length n. We input these premise-hypothesis pairs to RoBERTa:

$$S = [\text{<s>}\ \ P\ \ \text{</s>}\ \ H\ \ \ \text{</s>}\ ]\ ;\ T_r = \text{RoBERTa}(S)$$

Here, T_r denotes the context-aware representations of the premise and hypothesis.

5.3.2.2 Encoding premise and hypothesis

The encoder approach is inspired from [23]. We encode the Premise, $P = \{p_i\}_{i=1}^{m}$ and Hypothesis, $H = \{h_j\}_{j=1}^{n}$ using bidirectional LSTMs (BiLSTMs). We embed p_i and h_i into d_e dimensional vectors $\left[\mathbf{E}(p_1), ...\mathbf{E}(p_m)\right]$ and $\left[\mathbf{E}(h_1), ...\mathbf{E}(h_n)\right]$ using embedding matrix $\mathbf{E} \in \mathbb{R}^{d_e \times |V|}$, where $|V|$ is the Vocabulary size and $\mathbf{E}$ can be initialized with pretrained embeddings. We feed the premise-hypothesis pairs into BiLSTM encoders [96] to generate context-aware hidden states p^s and h^s.

$$p^s = \text{BiLSTM}(\mathbf{E(p)}, i)\ \ ;\ \ h^s = \text{BiLSTM}(\mathbf{E(h)}, i)$$

$$p^s \in \mathbb{R}^{m \times l_k} \text{and}\ h^s \in \mathbb{R}^{n \times l_k}$$

Here, l_k is the LSTM hidden state size. Following that we apply embedding dropout [70]

[2] Experiment with non-zero random initialization ref §5.4.3.

to enhance variation and prevent overfitting [296].

5.3.2.3 Premise and hypothesis attention module

To assess the contribution of external knowledge to the premise (and hypothesis), we utilize the Multi-Head dot-product attention [263] across knowledge representations and premise-hypothesis encoding. We calculate premise hypothesis relation values by normalizing relational connection embedding (R_{ijk}) with respect to column-axis (1), to obtain $R_{jk}^{prem} \in \mathbb{R}^{n \times k}$ which is the average premise relation for every hypothesis word.

$$R_{jk}^{prem} = \sum_{i=1}^{m} \frac{R_{ijk}}{m}$$

To apply dot product attention, we then reduce the dimension of the relation matrix to BiLSTM hidden state dimension, i.e., l_k.

$$R_{jk}^{r} = F_P^r(R_{jk}^{prem}) \quad \in \mathbb{R}^{n \times l_k}$$

where, F_P^r is a single layer neural network.

To highlight the importance of premise and its relations to hypothesis we utilise the premise attention head. The context-aware hypothesis hidden state h^s is used as queries, premise hidden state is used as keys and reduced premise hypothesis relation values are used as values. The attention function can be defined as follows:

$$\text{Attention}(h^s, p^s, R_{jk}^r) = \text{softmax}(\frac{h^s p^{sT}}{\sqrt{l}})R_{jk}^r$$

where, the multi-head attention is defined:

$$\begin{aligned} h_p^{att} &= \text{MH}(h^s, p^s, R_{jk}^r) \\ &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^o \end{aligned}$$

Here, $\text{head}_i = \text{Attention}(h^s W_i^q, p^s W_i^k, R_{jk}^r W_i^v)$ and W_i^q, W_i^k, and W_i^v are projection matrices and i is the number of attention heads. The output $h_p^{att} \in \mathbb{R}^{n \times l_k}$ is a context matrix that is attention-weighted according to the strength of the premise and its relationships to each of the hypothesis words. We also extract P^{att}, the premise multi-head attention attention weights. In hypothesis attention module, we use hypothesis attention head to highlight the importance of hypothesis and its relations to premise. Similar to the premise attention module, we calculate[3] $p_h^{att} \in \mathbb{R}^{m \times l_k}$, attention-weighted context matrix measuring the

[3]More details can be found in Appendix B.

importance of premise and relations to each of the hypothesis. We also extract H^{att}, the hypothesis multi-head attention attention weights.

5.3.2.4 Context aware external knowledge

ExBERT [68] uses a mixture model to weigh the balance of external relations and premise-hypothesis during inference. We construct attention-weighted external knowledge relations using Multi-head attention weights obtained in the attention modules.

$$P^{CE} = \sum_{k=1}^{h} P_{ij}^{att} R_{ijk} \;\; ; \;\; H^{CE} = \sum_{k=1}^{h} H_{ij}^{att} R_{ijk}$$

5.3.2.5 Composition layer

p^s encodes the individual word representations of the premise while p_h^{att} is the context representation of the premise aligned to the hypothesis. We can obtain word-level inference information for each word in the premise by composing them together with attention weights and context-aware external knowledge. We can do the same calculation for hypothesis, h^s and h_p^{att}:

$$p^m = G_P([p^s; p_h^{att}; p^s - p_h^{att}; p^s * p_h^{att}; \sum_{j=1}^{n} P_{ij}^{CE}])$$
$$h^m = G_H([h^s; h_p^{att}; h^s - h_p^{att}; h^s * h_p^{att}; \sum_{j=1}^{n} H_{ij}^{CE}])$$

Here, G_P and G_H are 2-layer neural networks with Dropout and ReLU activation [3] that compose the knowledge relations and premise-hypothesis contextual vectors into a unified knowledge aware context vector.

5.3.2.6 Pooling layer

The pooling layer creates fixed-length representations from the knowledge-aware premise and hypothesis context vectors.

$$p_{mean} = \text{MeanPool}(p^m) \; ; \; p_{max} = \text{MaxPool}(p^m)$$
$$h_{mean} = \text{MeanPool}(h^m) \; ; \; h_{max} = \text{MaxPool}(h^m)$$

5.3.2.7 Embedding mix-skip connection

To effectively integrate transformer embeddings with representations from premise and hypothesis, we introduce an Embedding mix-skip connection, where the embeddings

are concatenated and passed through a fully connected layer with a skip connection to transformer embeddings. Skip connections, introduced by [91], provides a shortcut to gradient flow and preserve the context between layers.

$$f = [p_{mean}, p_{max}, h_{mean}, h_{max}]$$

$$f' = T_r + F_c([T_r, f])$$

Here, F_c is a two-layer neural network with dropout and ReLU activation. Finally, f' is passed through a classification layer to obtain the inference class.

5.4 Experiment and Analysis

Our experiments study the following questions.

RQ1: Is our proposed model competent in using external knowledge sources effectively to enhance performance across INFOTABS evaluations sets?

RQ2: How effective is our approach in settings with little supervision? How much supervision is necessary to outperform benchmark models?

RQ3: (a) Which reasoning types is our proposed model most effective at boosting? (b) Is our approach equally effective across all domains, that is, across all table categories?

RQ4: How does the model component choices impact performance? (a) To what extent are skip connections, (b) knowledge embeddings, (c) additional MNLI [280] pre-finetuning, and (d) a bigger pre-trained model beneficial?

5.4.1 Experimental Setup

Here, we discuss the datasets, external knowledge sources, and the models used in the experiments.

5.4.1.1 Datasets

We use INFOTABS, a tabular Language inference dataset introduced by [84] for all our experiments. The dataset is diverse in categories and keys and requires background knowledge and semantic understanding of the text. Examples in INFOTABS are labeled with three types of inference: entailment, neutrality, and contradiction, based on their relation with premise tables. Along with the standard development set and test set (dubbed α_1), the dataset includes two adversarial test sets: a contrast set dubbed α_2 that is lexically similar

to α_1 but contains fewer hypotheses, and a zero-shot set dubbed α_3 that contains long tables from various domains with little key overlap with the training set.

5.4.1.2 Table representation

To represent tables, we utilize [182] *Better Paragraph Representation* (BPR) technique in conjunction with *Distracting Row Removal* (DRR). The BPR technique turns its rows into sentences using a universal template, enabling it to be used as the input for a BERT-style model. We utilize the DRR approach to reduce the premise table by identifying the most relevant premise sentence. For finding the most relevant rows, we use cosine similarity over fastText embeddings [15] and word alignment with the specified hypothesis. We select the top four aligned table rows from each premise table with hypotheses.

5.4.1.3 Knowledge sources

We utilize ConceptNet, as introduced by [243] to extract external commonsense knowledge to create relational occurrences. We notice that 85% of premise-hypothesis pairings contain at least one relationship in the ConceptNet database. To supplement the coverage, we also use Wordnet [170], to extract additional lexical word relations, namely *Synonyms*, *Antonyms*, *Hypernyms*, *Hyponyms* and *Co-Hyponyms*. After combining the two knowledge databases and removing duplicates, the number of non-zero relational connection pairings increases to 90%. We create an English directional single word relations dataset by merging ConceptNet and Wordnet. The combined KG source contains 11.2 million relation triples. For example in the Table 5.1, the relational occurrence { "coast" $\leftarrow$ "California"} extracted from Conceptnet, provide the necessary world knowledge required for correct inference.

5.4.1.4 Word embeddings

We utilize pre-learned word embeddings to initialize the BiLSTM encoders. The premise and hypothesis words are embedded in 300-dimensional vectors using Glove embeddings introduced by [198]. We also investigate fastText embeddings for representation, but it has only 77.4 % coverage of all tokens. Glove is a collection of 400,000-word embeddings learned using the Wikipedia, Common crawl, and Twitter datasets. We realize that the

GloVe vocabulary covers 85.6% of the terms in INFOTABS dataset.[4]

5.4.1.5 Models

To evaluate we compare our model with INFOTABS [84] and Knowledge-INFOTABS [182] baselines, specifically we employ the following methods:

- **RoBERTa**. The original RoBERTa baseline of INFOTABS. We append and encode premise and hypothesis pairs with BPR with DRR representation and generate an inference label with the RoBERTa classification head.
- **KG Explicit.** Knowledge-INFOTABS introduced this baseline. The baseline uses the same RoBERTa classifier as the INFOTABS, except that the premise end is augmented with extracted premise row key definitions from Wordnet and Wikipedia sources before encoding and classifying using RoBERTa. Additionally, prior to appending, the method employs key sense disambiguation to assure that only relevant hypothesis context-related definitions are added. For example, for a table with category *"Person"* and key *"Spouse"*, the definition of *"Spouse"* from Wikipedia, i.e., *"Spouse is defined as a spouse is a significant other in a marriage, civil union, or common-law marriage."* is appended as a suffix.
- **Tok-KTrans**. We utilize Wordnet to expand premise hypothesis pairs with word relations in Tokens added transformers before encoding and classifying using RoBERTa. We extend the tokenizer by including relational tokens and appending the relationships with the following format - $\{$<*KNW*> *[premise_word*$_1$ *: hypothesis_word*$_1$ *;* <*relation*$_1$> *] [premise_word*$_2$ *: hypothesis_word*$_2$ *;* <*relation*$_2$>*] ...* $\}$. For example, The table *Jallikattu* contains a key ***Mixed Gender*** with a value *NO*. The hypothesis, *Jallikattu is a single* ***sex*** *sport* contradicts the premise table. We append the relation $\{$<*KNW*> *[gender : sex ;* <*SYN*> *]*$\}$ as suffix to input prior to the RoBERTa classification.
- **Trans-KBLSTM.** This is our proposed model as described in the §5.3. For details on model training and hyper-parameters, refer to Appendix B.

[4]Due to limited supervision, we found that freezing word embedding during the BiLSTM training is beneficial. For the remaining unseen tokens, we initialized with zero vectors.

5.4.2 Results and Analysis

This section summarizes our findings concerning the research questions.

5.4.2.1 Full supervision setting

To assess the effectiveness of our method Trans-KBLSTM (i.e. RQ1), we train baseline and our model Trans-KBLSTM with 100% of training data. Table 5.2 shows the performance (accuracy) for all models. We observe that Trans-KBLSTM outperform[5] all other baselines. On development, α_1, and α_3 Trans-KBLSTM outperform 0.75 - 0.95 % with 100% training data.

5.4.2.2 Limited supervision setting

To ensure that our model works effectively in low-resource scenarios (i.e., RQ2), we analyze models trained under limited supervision. We randomly sampled {1, 2, 3, 5, 10, 15, 20, 25, 30, 50, and 100} data in an incremental method [6]. We experimented three times using random seeds for sampling/training to account for sample variability.

Figure 5.2 shows the accuracy for all models. We observe a huge performance improvement with Tran-KBLSTM over other baseline models for low data regimes. All improvements are statistically significant with Student's t-test $p < 0.05$ except dev results with 3% and 5%. For precise numbers and standard deviation plots, see Appendix B. Additionally, as the training supervision increases, the performance margin across models narrows. This improvement can be attributed to the fact that the model's reasoning ability increases when more training data is added, resulting in more accurate predictions without explicitly necessitating external knowledge addition. As a result, adding external knowledge may not be as beneficial if there is adequate supervision.

5.4.2.3 Reasoning analysis

To investigate the reasoning behind a model's prediction (i.e., RQ3(a)), INFOTABS adapted the set of reasoning categories from GLUE [270] for tabular premises. Thus, we also analyze performance across several reasoning types on the development set of INFOTABS. We utilized the reasoning annotated instances from INFOTABS for our analysis. Figure 5.3

[5] reaches maximum in 6-7 epochs while [182] takes 14-15 epochs

[6] Higher % include all instances from lower %, i.e. a 20% includes all instances from a 10% samples.

(1%) and Figure 5.4 (3%) show the performance across various reasoning types on the development set for 1% and 3% of INFOTABS development set. Trans-KBLSTM model shows improvements in several reasoning types including *"Lexical"*, *"Multi-Row"*, and *"KCS"*.

- *Lexical Reasoning* involves inferencing through words independent of context, where the word falls. Since we add relational connections between words which include synonyms, antonyms, etc. lexical reasoning ability of the model enhances. For example, in the table *"Chibuku Shake"*, the key *"Ingredients"* contains *"Sorghum"* and *"Maize"* while the hypothesis requires us to infer about *Corn* as an ingredient in the Chibuku shake. The relation {"corn" $\xleftarrow{Synonym}$ "Maize"} helps the model in making the correct prediction. For details refer Appendix B.
- *Multi-Row Reasoning* involves making an inference using multiple rows of the table. When the reasoning involves multiple rows, the model needs to extract the relevant rows and rightly focus on selected related connected phrases. The relational connections that we propose between premise and hypothesis tokens establish these extractions and connections and thus enhancing the multi-row reasoning ability of the Trans-KBLSTM model. For example in a "Person" table relations such as { "born" $\xleftrightarrow{RelatedTo}$ "young" ; "born" $\xleftrightarrow{RelatedTo}$ "child" ; "child" $\xleftrightarrow{RelatedTo}$ "age" ; "year active" $\xleftrightarrow{Co-Hyponym}$ "child" } help in connected both the born, child and year active keys with the concern hypothesis. For details refer Appendix B.
- *Knowledge and Common Sense Reasoning.* This reasoning is related to the World Knowledge and Common Sense category from GLUE-Benchmark [270], which is quoted as "...the entailment rests not only on correct disambiguation of the sentences, but also, application of extra knowledge, whether factual knowledge about world affairs or more commonsense knowledge about word meanings or social or physical dynamics." Knowledge databases like ConceptNet contain many knowledge relations capable of enhancing these reasoning type. For example, in a "Country" table relations such as { "kingdom" $\xleftrightarrow{IsA}$ "monarchy" ; "democracy" $\xleftrightarrow{RelatedTo}$ "Government" } add additional information necessary for inference. For details refer Appendix B.

Improvement across Inference Labels: In our analysis, we observe a performance improvement across the Entailment and Neutral labels, but only a negligible increase, for example, in instances labeled with the Contradiction label. Contradictory label prediction

requires noise-free, contextually relevant knowledge to ascertain the negation. External knowledge addition with minimal noise can lead to the predicted Neutral or Entailment label. Additional ways for relational connection trimming may be explored in future.

5.4.3 Ablation Study

We perform ablation studies (i.e., RQ4) to understand the importance of individual model components further. The ablation study was conducted to ascertain the significance of (a) Trans-KBLSTM Skip Connection, (b) Knowledge Relations, (c) Implicit KG addition via. MNLI pre-training (Embeddings), and (d) Transformer Model Param Size. (e) Independent Component training.

5.4.3.1 Effect of skip connections

We study the significance of embedding skip connection and the knowledge relations (i.e., RQ4(a,b)). The knowledge relations are initialized with random vectors to examine model performance variations.

Table 5.3 shows the Trans-KBLSTM performance with several ablations. We observe that adding knowledge and the introduction of skip connection improve the model performance. The addition of knowledge to the model improves the performance on Dev, $\alpha 1$, and $\alpha 2$ sets. The inclusion of knowledge improves performance the most for Development, α_2, and α_3 sets, whereas the addition of skip connection improves performance substantially in α_1 set. The performance improvement in α_3 set demonstrates that using external information benefits zero-shot settings (i.e., cross-domain transfer learning). The improved performance by the addition of skip connection demonstrates that effective knowledge integration significantly impacts model performance.

5.4.3.2 Implicit knowledge addition

We examine the effect of implicit knowledge addition (i.e., RQ4(b)) in Trans-KGLSTM model. Thus, similar to the KG Implicit baseline of Knowledge-INFOTABS [182], we supplement implicit knowledge using the MNLI via data augmentation. To ensure a fair comparison, we compare the two Trans-KBLSTM RoBERTa-based classifiers, one with and the other without MNLI data pre-training. The performance with MNLI pre-training is reported in Figure 5.5.

We observe an improvement in performance for all percentages of train data after pre-training using MNLI data. Pre-training enables the model to acquire domain-specific information, hence enhancing its performance. There is a more significant gain in performance for non-pre-trained than for MNLI pre-trained models, suggesting that external information addition is more beneficial for models without any implicit knowledge. In comparison, our approach uses relational connections to augment the model's knowledge in the phase, final training avoiding the computational, time, and economic cost of large MNLI pre-training.

5.4.3.3 Effect of transformer size

We substitute RoBERTa$_{\text{LARGE}}$ with RoBERTa$_{\text{BASE}}$ to study the effect of transformer size on performance (i.e. RQ4 (d)) of INFOTABS test sets. We pre-train both the transformers model using the MultiNLI dataset for all percentages. The performance is depicted in Figure 5.6. We see an increase in performance as the model's size increases, especially for external knowledge addition, i.e., Trans-KBLSTM model.

5.4.3.4 Independent training

We examine the effect of training transformer and KBLSTM components independently. For independent training, we first train RoBERTa$_{\text{LARGE}}$ transformer model on INFOTABS. Then we utilize these weights to initialize the transformer component of Trans-KBLSTM. Finally, we trained the KBLSTM component of Trans-KBLSTM on INFOTABS while keeping these pre-trained transformer weights frozen (constant). Table 5.4 shows the results of training Trans-KBLSTM with different regimes. We observe that training the components together shows a more significant improvement in performance than training the KBLSTM component independently. Joint training of transformer and KBLSTM generates representations in the same embedding space, enhancing external knowledge integration.

5.5 Conclusion and Future Work

In this chapter, we introduce Trans-KBLSTM, a novel architecture to integrate external knowledge into tabular NLI models. Trans-KBLSTM is shown to improve reasoning on the INFOTABS dataset. The performance advantage is particularly pronounced in low-data regimes. The reasoning study demonstrates that the model enhances lexical, numerical,

and multiple-row reasoning. Ablation experiments demonstrate the critical nature of each component in the model's design. We believe that our findings will be valuable to researchers working on the integration of external knowledge into deep learning architectures. Performance of the proposed architecture on more datasets can be explored in future studies. Looking forward, the application of this architecture to other NLP tasks that can benefit from external knowledge enhanced relational connections between sentence pairs, such as question answering and dialogue understanding.

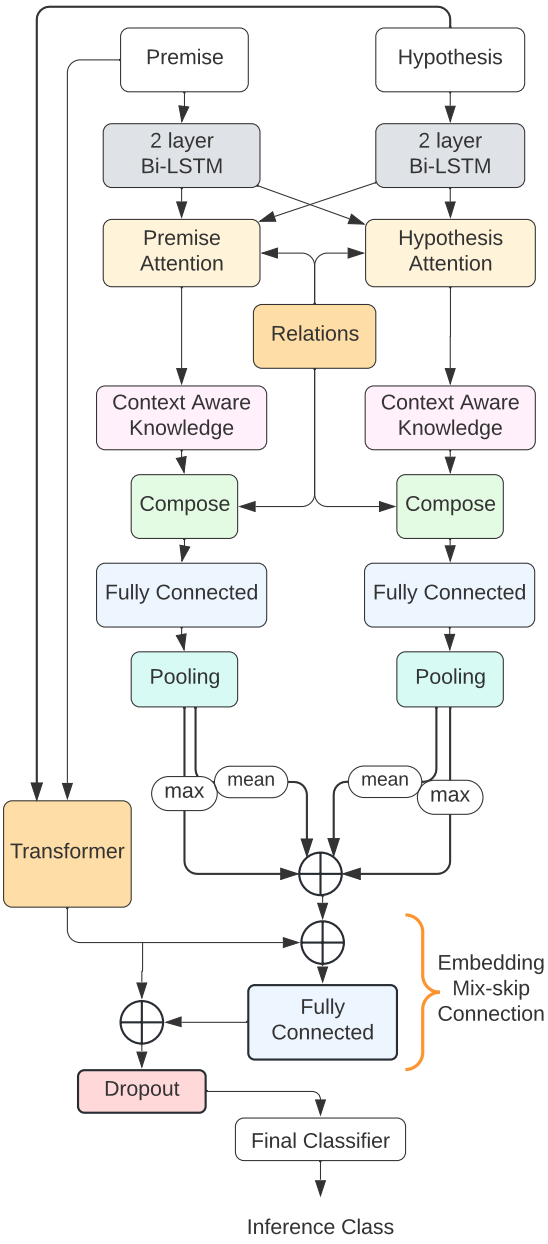


Figure 5.1: High level flowchart of Trans-KBLSTM.

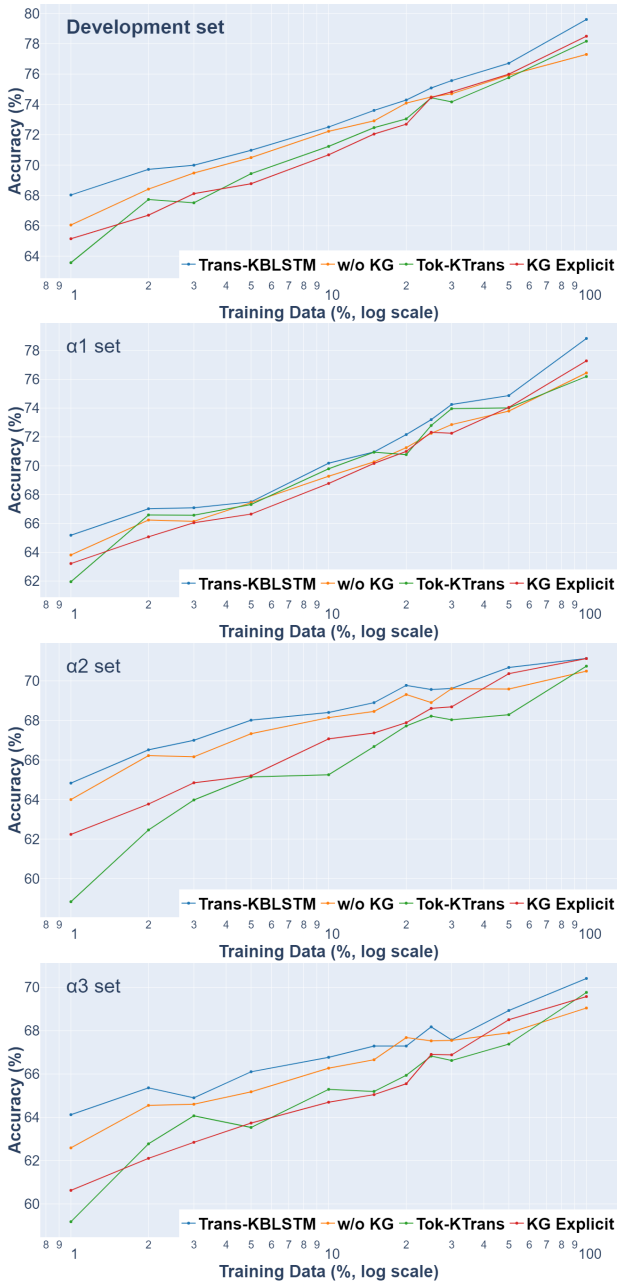


Figure 5.2: Performance in terms of accuracy in limited supervision setting. **w/o KG** represent RoBERTa INFOTABS [84] baseline, **KG Explicit** represent Knowledge-INFOTABS [182] baseline, **Tok-KTrans** is the token appended transformers and **Trans-KBLSTM** represent our proposed model. Reported results are average over 3 random seed runs with average standard deviation of 0.233 (w/o KG), 0.49 (KG Explicit), 0.50 (Tok-KTrans) and 0.30 (Trans-KBLSTM). All the improvements are statistically significant with Student's t-test $p < 0.05$ of one-tailed Student t-test.

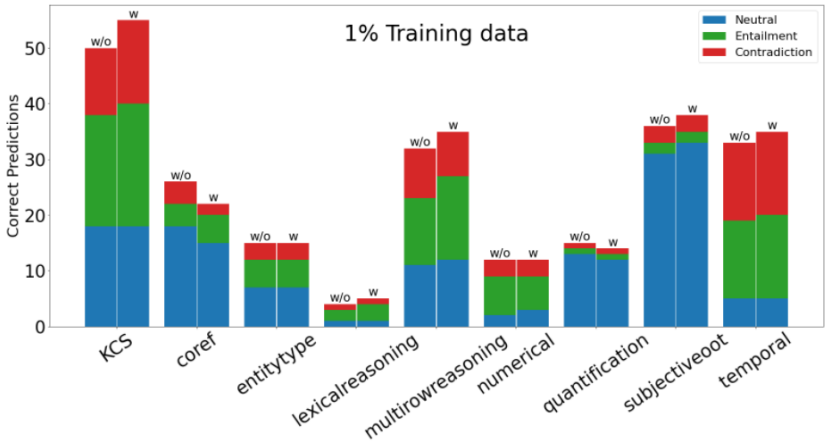


Figure 5.3: Number of correct model predictions across various reasoning types. **w/o** represents without knowledge (KG) i.e. original RoBERTa models and **w** represents Trans-KBLSTM model with explicitly added relational connection knowledge (KG).

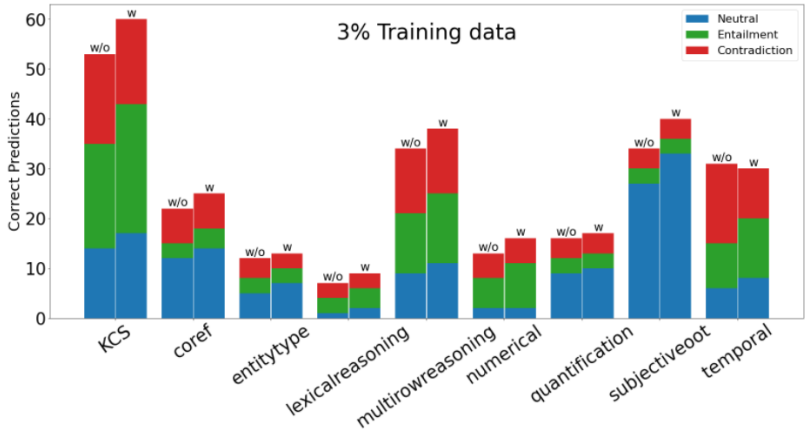


Figure 5.4: Number of correct model predictions across various reasoning types. **w/o** represents without knowledge (KG) i.e. original RoBERTa models and **w** represents Trans-KBLSTM model with explicitly added relational connection knowledge (KG).

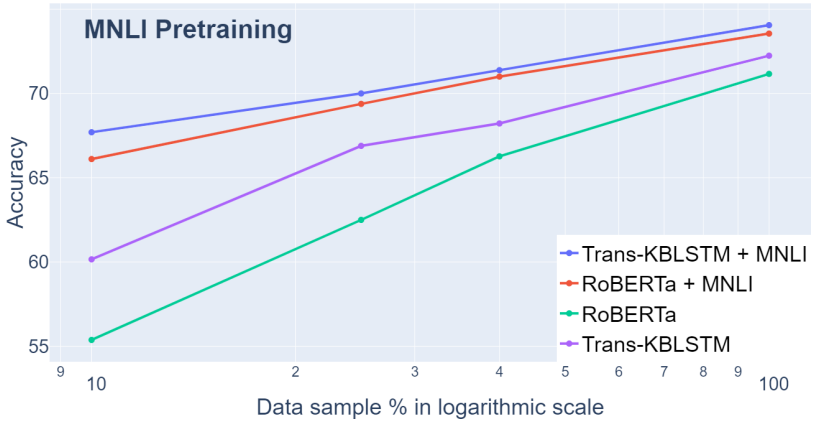


Figure 5.5: Performance improvement with MNLI pre-training across various models.

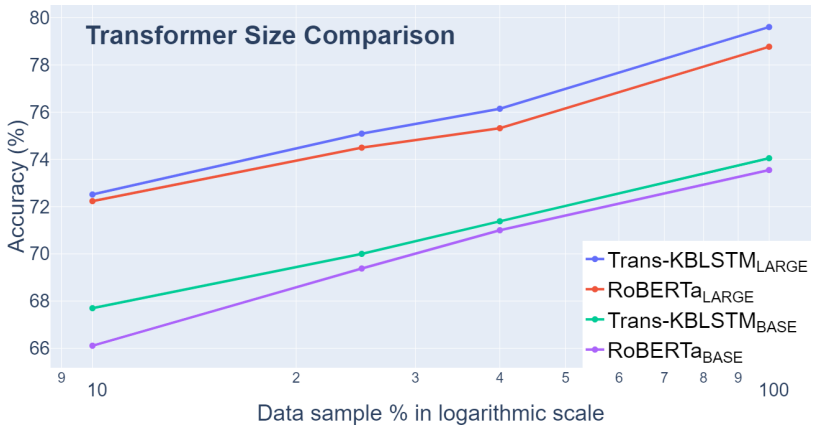


Figure 5.6: Improvement in model performance across varying models sizes.

Table 5.1: An INFOTABS example demonstrating the need of knowledge augmentation. Predicting the Gold label requires broad understanding of *California* is located on the *Coast*. In the table, for each row the first column represents the keys (unique identifiers) and the second column represents their corresponding values (attributes).

James Hetfield	
Birth Name	James Alan Hetfield
Born	Aug. 3, 1963(age 58), California, U.S.
Genres	Heavy metal, thrash metal, hard rock
Occupation(s)	Musician, Singer
Instruments	Vocals, Guitar
Years active	1978-present
Labels	Warner Bros, Elektra, MegaForce
Hypothesis	James Hetfield was born on the west coast of the USA.
Focused Relation	coast $\xleftarrow{AtLocation}$ california
Human	Entailment
RoBERTa	Neutral
Trans-KBLSTM	Entailment

Table 5.2: Performance in terms of accuracy with full supervision. **w/o Knoweldge** represent RoBERTa INFOTABS [84] baseline, **KG Explicit** represent Knowledge-INFOTABS [182] baseline, **Tok-KTrans** is the token appended transformers and **Trans-KBLSTM** represent our proposed model. Reported number are average over three random seeds with standard deviation of 0.27 (w/o KG), 0.69 (Tok-KTrans), 0.23 (KG Explicit) and 0.36 (Trans-KBLSTM). All improvements are statistically significant with Student's t-test $p < 0.05$ except α_2 with KG Explicit.

Model	Dev	$\alpha 1$	$\alpha 2$	$\alpha 3$
w/o Knowledge	77.30	76.44	70.49	69.05
Tok-KTrans	78.17	76.19	70.75	69.77
KG Explicit	78.97	77.84	71.13	69.58
Trans-KBLSTM	**79.92**	**79.62**	**72.10**	**70.21**

Table 5.3: Ablation study performance on stratified 1% split of dataset. We systematically eliminate model components in order to evaluate the performance improvement.

Ablations	Dev	α_1	α_2	α_3
Trans-KBLSTM	**67.55**	**65.16**	**64.00**	**63.38**
- Skip Connect	65.72	62.83	60.00	61.55
- KB	60.44	61.88	56.94	55.55
- (KB + Skip Connect)	60.11	61.50	55.94	57.38

Table 5.4: Joint/Independent training performance on INFOTABS dataset. First row shows results of training only RoBERTa$_{\text{LARGE}}$ model without knowledge. Second row shows results of training KBLSTM independently after freezing RoBERTa$_{\text{LARGE}}$ parameters. Third row shows the results of our proposed approach i.e. Joint-training of RoBERTa$_{\text{LARGE}}$ and KBLSTM.

Ablations	Dev	α_1	α_2	α_3
RoBERTa$_{\text{LARGE}}$	77.30	76.44	70.49	69.05
+ KBLSTM (Independent)	79.22	78.38	71.00	69.22
+ KBLSTM (Joint Train)	**79.92**	**79.62**	**72.10**	**70.21**

CHAPTER 6

SYSTEMATIC TABULAR PROBES

Adapted from V. Gupta, R. Bhat, A. Ghosal, M. Shrivastava, M. Singh, and V. Srikumar. *Is my model using the right evidence? Systematic probes for examining evidence-based tabular reasoning.* Trans. Assoc. Comput. Linguist. (TACL), 10 (2022), pp. 659–679.

One strategy for tabular reasoning tasks relies on the successes of contextualized representations [51, 158] for the sentential version of the problem. Tables are flattened into artificial sentences using heuristics to be processed by these models, as described in Chapter 3. Surprisingly, even this naïve strategy leads to high predictive accuracy, as shown not only in Chapter 3 but also by related lines of recent work [58, 291].

In this chapter, we ask: *Do these seemingly accurate models for tabular inference effectively use and reason about their semi-structured inputs?* While "reasoning" can take varied forms, a model that claims to do so should at least ground its outputs on the evidence provided in its inputs. Concretely, we argue that such a model should (a) be self-consistent in its predictions across controlled variants of the input, (b) use the evidence presented to it, and the right parts thereof, and, (c) avoid being biased *against* the given evidence by knowledge encoded in the pre-trained embeddings.

6.1 Contributions

Corresponding to these three properties, we identify three dimensions to evaluate a tabular NLI system: robustness to annotation artifacts, relevant evidence selection, and robustness to counterfactual changes. We design systematic probes that exploit the semi-structured nature of the premises. This allows us to semi-automatically construct the probes and to unambiguously define the corresponding expected model response. These probes either introduce controlled edits to the premise or the hypothesis, or to both, thereby also creating counterfactual examples. Experiments reveal that despite seemingly high test set accuracy, a model based on RoBERTa [158], a good representative of BERT derivative

models, is far from being reliable. Not only does it ignore relevant evidence from its inputs, it also relies excessively on annotation artifacts, in particular the sentence structure of the hypothesis, and pre-trained knowledge in the embeddings. Finally, we found that attempts to inoculate the model [155] along these dimensions degrades its overall performance.[1]

This work is published at TACL 2022, and presented at ACL 2022 as [83]. Additionally, we also released a interactive annotation platform for generating effective tabular perturbations, which got published in EMNLP 2021 Demo track as [104]. TabPert facilitates this by generation of such counterfactual data for assessing model tabular reasoning issues. TabPert allows the user to update a table, change the hypothesis, change the labels, and highlight rows that are important for hypothesis classification. TabPert also details the technique used to automatically produce the table, as well as the strategies employed to generate the challenging hypothesis. These counterfactual tables and hypotheses, as well as the metadata, is then used to explore the existing model's shortcomings methodically and quantitatively.

6.2 Background

Unlike unstructured data, where creating challenge datasets may be more difficult [79, 172, 221], we can analyze semi-structured data more effectively. Although connected with the title, the rows in the table are still independent, linguistically and otherwise. Thus, controlled experiments are easier to design and study. For example, the analysis done for evidence selection via multiple table perturbation operations such as row deletion and insertion is possible mainly due to the tabular nature of the data. Such granularity and component-independence is generally absent for raw text at the token, sentence and even paragraph level. As a result, designing suitable probes with sufficient coverage can be a challenging task, and can require more manual effort.

Additionally, probes defined on one tabular dataset (INFOTABS in our case) can be easily ported to other tabular datasets such as WikiTableQA [195], TabFact [26], HybridQA [27, 189, 297], OpenTableQA [24], ToTTo [194], Turing Tables [293], LogicTable [25]. Moreover, such probes can be used to better understand the behavior of various tabular reasoning

[1]The dataset and the scripts used for our analysis are available at https://tabprobe.github.io.

models [77, 94, 101, 177, 204, 291].

6.2.1 Interpretability for NLI Model

For classification tasks such as NLI, correct predictions do not always mean that the underlying model is employing correct reasoning. More work is needed to make models interpretable, either through explanations or by pointing to the evidence that is used for predictions [52, 63, 95, 105, 187, 192, 215, 223, 234, 279]. Many recent shared tasks on reasoning over semi-structured tabular data (such as SemEval 2021 Task 9 [226] and FEVEROUS [5]) have highlighted the importance of, and the challenges associated with, evidence extraction for claim verification.

Finally, NLI models should be tested on multiple test sets in adversarial settings [78, 103, 155, 167, 179, 184, 218, 219, 220, 307] focusing on particular properties or aspects of reasoning, such as perturbed premises for evidence selection, zero-shot transfer (α_3), counterfactual premises or alternate facts, and contrasting hypotheses via perturbation (α_2). Such behavioral probing by evaluating on multiple test-only benchmarks and controlled probes is essential to better understand both the abilities and the weaknesses of pre-trained language models.

6.3 Preliminaries: Tabular NLI

Tabular natural language inference is a task similar to standard NLI in that it examines if a natural language hypothesis can be derived from the given premise. Unlike standard NLI, where the evidence is presented in the form of sentences, the premises in tabular NLI are semi-structured tables that may contain both text and data.

6.3.1 Dataset

Recently, datasets such as TabFact [26] and INFOTABS [84], and also shared tasks such as SemEval 2021 Task 9 [226] and FEVEROUS [5], have sparked interest in tabular NLI research. In this study, we use the INFOTABS dataset for our investigations.

INFOTABS consists of 23,738 premise-hypothesis pairs, whose premises are based on Wikipedia infoboxes. Unlike TabFact, which only contains ENTAIL and CONTRADICT hypotheses, INFOTABS also includes NEUTRAL ones. Table 6.1 shows an example table from the dataset with four hypotheses, which will be our running example.

The dataset contains 2,540 distinct infoboxes representing a variety of domains. All hypotheses were written and labeled by MTurk workers. The tables contain a *title* and two columns, as shown in the example. Since each row takes the form of a key-value pair, we will refer to the elements in the left column as the *keys*, and the right column provides the corresponding *values*.

In addition to the usual train and development sets, INFOTABS includes three test sets, α_1, α_2 and α_3. The α_1 set represents a standard test set that is both topically and lexically similar to the training data. In the α_2 set, hypotheses are designed to be lexically adversarial, and the α_3 tables are drawn from topics not present in the training set. We will use all three test sets for our analysis.

6.3.2 Models over Tabular Premises

Unlike standard NLI, which can use off-the-shelf pre-trained contextualized embeddings, the semi-structured nature of premises in tabular NLI necessitates a different modeling approach.

Following [26], tabular premises are flattened into token sequences that fit the input interface of such models. While different flattening strategies exist in the literature, we adopt the *Table as a Paragraph* strategy of [84], where each row is converted to a sentence of the form "The `key` of `title` is `value`". This seemingly naïve strategy, with RoBERTa-large embeddings (RoBERTa$_L$ henceforth), achieved the highest accuracy in the original work, shown in Table 6.2.[2] The table also shows the hypothesis-only baseline [87, 202] and human agreement on the labels.[3]

To study the stability of the models to variations in the training data, we performed 5-fold cross validation (5xCV). An average cross validation accuracy of 73.53% with a standard deviation of 2.73% was observed on the training set which is close to the performance on the α_1 test set (74.88%). In addition, we also evaluated performance on the

[2]Other flattening strategies have similar performances [84].

[3]Preliminary experiments on the development set showed that RoBERTa$_L$ outperformed other pre-trained embeddings. We found that BERT$_B$, RoBERTa$_B$, BERT$_L$, ALBERT$_B$ and ALBERT$_L$ reached development set accuracies of 63.0%, 67.23%, 69.34%, 70.44% and 70.88%, respectively. While we have not replicated our experiments on these other models due to a prohibitively high computational cost, we expect the conclusions to carry over to these other models as well.

development and test sets. The penultimate row of Table 6.2 presents the performance for the model trained on the entire training data, while the last row presents the performance of the 5xCV models. The results demonstrate that model performance is reasonably stable to variations in the training set.

6.4 Reasoning: An Illusion?

Given the surprisingly high accuracies in Table 6.2, especially on the α_1 test dataset, can we conclude that the RoBERTa-based model reasons effectively about the evidence in the tabular input to make its inference? That is, does it arrive at its answer via a sound logical process that takes into account all available evidence along with common sense knowledge? Merely achieving high accuracy is not sufficient evidence of reasoning: the model may arrive at the right answer for the wrong reasons leading to improper and inadequate generalization over unseen data. This observation is in line with the recent work pointing out that the high-capacity models we use may be relying on spurious correlations [202].

"Reasoning" is a multi-faceted phenomenon, and fully characterizing it is beyond the scope of this work. However, we can probe for the *absence* of evidence-grounded reasoning via model responses to carefully constructed inputs and their variants. The guiding premise for this work is:

> *Any "evidence-based reasoning" system should demonstrate expected, predictable behavior in response to controlled changes to its inputs.*

In other words, "reasoning failures" can be identified by checking if a model deviates from expected behavior in response to controlled changes to inputs. We note that this strategy has been either explicitly or implicitly employed in several lines of recent work [72, 221]. In this work, we instantiate the above strategy along three specific dimensions, briefly introduced here using the running example in Table 6.1. Each dimension is used to define several concrete probes that subsequent sections detail.

6.4.1 Avoiding Annotation Artifacts

A model should not rely on spurious lexical correlations. In general, it should not be able to infer the label using only the hypothesis. Lexical differences in closely related hypotheses should produce predictable changes in the inferred label. For example, in the hypothesis H2 of Table 6.1 if the token "end" is replaced with "start", the model prediction

should change from CONTRADICT to ENTAIL.

6.4.2 Evidence Selection

A model should use the correct evidence in the premise for determining the hypothesis label. For example, ascertaining that the hypothesis H1 is entailed requires the *Genre* and *Length* rows of Table 6.1. When a relevant row is removed from a table, a model that predicts the ENTAIL or the CONTRADICT label should predict the NEUTRAL label. When an irrelevant row is removed, it should not change its prediction from ENTAIL to NEUTRAL or vice versa.

6.4.3 Robustness to Counterfactual Changes

A model's prediction should be *grounded* in the provided information even if it contradicts the real world, i.e., to counterfactual information. For example, if the month of the *Released* date changed to "December", then the model should change the label of H2 in Table 6.1 to ENTAIL from CONTRADICT. Since this information about release date contradicts the real world, the model cannot rely on its pre-trained knowledge, say from Wikipedia. For the model to predict the label correctly, it needs to reason with the information in the table as the primary evidence. Although the importance of pre-trained knowledge cannot be overlooked, it must not be at the expense of primary evidence.

Further, there are certain pieces of information in the premise (irrelevant to the hypothesis) which do not impact the outcome, making the outcome *invariant* to these changes. For example, deleting irrelevant rows from the premise should not change the model's predicted label. Contrary to this is the relevant information ("evidence") in the premise. Changing these pieces of information should vary the outcome in a predictable manner, making the model *covariant* with these changes. For example, deleting relevant evidence rows should change the model's predicted label to NEUTRAL.

The three dimensions above are not limited to tabular inference. They can be extended to other NLP tasks, such as reading comprehension as well as the standard sentential NLI. However, directly checking for such properties there would require a lot of labeled data—a big practical impediment. Fortunately, in the case of tabular inference, the (in-/co-)variants associated with these dimensions allow controlled and semi-automatic edits to the inputs leading to predictable variation of the expected output. This insight underlies

the design of probes using which we examine the robustness of the reasoning employed by a model performing tabular inference. As we will see in the following sections, highly effective and precise probes can be designed without extensive annotation.

6.5 Probing Annotation Artifacts

Can a model make inference about a hypothesis without a premise? It is natural to answer in the negative in general (Of course, certain hypotheses may admit strong priors, e.g., tautologies.). Preliminary experiments by [84] on INFOTABS, however, reveal that a model trained just on hypotheses performs surprisingly well on the test data. This phenomenon, an inductive bias entirely predicated on the hypotheses, is called *hypothesis bias*. Models for other NLI tasks have been similarly shown to exhibit hypothesis bias, whereby the models learn to rely on spurious correlations between patterns in the hypotheses and corresponding labels [75, 87, 202]. For example, negations are observed to be highly correlated with contradictions [187].

To better characterize a model's reliance on such artifacts, we perform controlled edits to hypotheses without altering associated premises. Unlike the α_2 set, which includes minor changes to function words, we aim to create more sophisticated changes by altering content expressions or noun phrases in a hypothesis. Two possible scenarios arise where a hypothesis alteration, without a change in the premise, either (a) leads to a change in the label (i.e., the label covaries with the variation in the hypothesis), or (b) does not induce a label change (i.e., the label is invariant to the variation in the hypothesis).

In INFOTABS, a set of reasoning categories are identified to characterize the relationship between a tabular premise and a hypothesis. We use a subset of these, listed below, to perform controlled changes in the hypotheses: (a.) **Named Entities**: entities such as *Person, Location, Organisation*, (b.) **Nominal modifiers**: nominal phrases or clauses, (c.) **Negation**: markers such as *no, not*, (d.) **Numerical Values**: numeric expressions such as *weights, percentages, areas*, (e.)**Temporal Values**: Date and Time, and (f.) **Quantification**: quantifiers such as *most, many, every*.

Although we can easily track these expressions in a hypothesis using tools like entity recognizers and parsers, it is non-trivial to automatically modify them with a predictable change on the hypothesis label. For example, some label changes can only be controlled

if the target expression in the hypothesis is correctly aligned with the facts in the premise. Such cases include CONTRADICT to ENTAIL, and NEUTRAL to CONTRADICT or ENTAIL, which are difficult without extensive expression-level annotations. Nonetheless, in several cases, label changes can be deterministically known even with imprecise changes in the hypothesis. For example, we can convert a hypothesis from ENTAIL to CONTRADICT by replacing a named entity in the hypothesis with a random entity of the same type.

Hence we follow the following strategy: (a) We avoid perturbations involving the NEUTRAL label altogether, as they often need changes in the premise (table) as well. (b) We generate all label-preserving and some label-flipping transformations automatically using the approach described below. (c) We annotate the CONTRADICT to ENTAIL label-flipping perturbations manually.

6.5.1 Automatic Generation of Label Preserving Transformations

To automatically perturb hypotheses, we leverage the syntactic structure of a hypothesis and the monotonicity properties of function words like prepositions. First, we perform syntactic analysis of a hypothesis to identify named entities and their relations to title expressions via dependency paths.[4] Then, based on the entity type, we either substitute or modify them. Named entities such as person names and locations are substituted with entities of the same type. Expressions containing numbers are modified using the monotonicity property of the prepositions (or other function words) governing them in their corresponding syntactic trees.

Given the monotonicity property of a preposition (see Table 6.3), we modify its governing numerical expression in a hypothesis in the same order to preserve the hypothesis label. Consider the hypothesis H5 refered below as Figure 6.1 which contains a preposition (*over*) with upward monotonicity. Because of upward monotonicity, we can increase the number of hours in H5 without altering the label.

Manual annotation of label-flipping transformations: Note that in the above example, modifying the numerical expression in the reverse direction (e.g., decreasing the number of hours) does not guarantee a label flip. We need to know the premise to be accurate. During the experiments, we observed that a large step (half/twice the actual number)

[4]We used spaCy v2.3.2 for the syntactic analysis.

suffices in most cases. We used this heuristic and manually curated the erroneous cases. Additionally, all the cases of CONTRADICT to ENTAIL label-flipping perturbations were annotated manually.[5]

We generated 2,891 perturbed examples from the α_1 set with 1,203 instances preserving the label and 1,688 instances flipping it. We also generated 11,550 examples from the *Train* set, with 4,275 preserving and 7,275 flipping the label. Some example perturbations using different types of expressions are listed in Table 6.4. It should be noted that there may not be a one-to-one correspondence between the gold and perturbed examples, as a hypothesis may be perturbed numerous times or not at all. As a result, in order for the results to be comparable, a single perturbed example must be sampled for each gold example: we sampled 967 from the α_1 set and 4,274 from the *Train* set.

6.5.2 Results and Analysis

We tested the hypothesis-only and full models (both trained on the original *Train* set) on the perturbed examples, without subsequent fine-tuning on the perturbed examples.[6] The results are presented in Table 6.5, with each cell representing the average accuracy and standard deviation (subscript) across 100 samplings, with 80% of the data selected at random in each sampling.

We note that the performance degrades substantially in both label-preserved and flipped settings when the model is trained on just the hypotheses. When labels are flipped after perturbations, the decrease in performance (averaged across both models) is about 25% and 61% points, on the α_1 set and *Train* set respectively. However, for the full model, perturbations that retain the hypothesis label have little effect on model performance.

The contrast in the performance drop between the label-preserved and label-flipped cases suggests that changes to the content expressions have little effect on the model's original predictions. Interestingly, the predictions are invariant to changes to functions words as well, as per results on α_2 in [84]. This suggests that the model might be more prone to changes to the template or structure of a hypothesis than its lexical makeup. Consequently, a model that relies on correlations between the hypothesis structure and

[5] Annotation done by an expert well versed in the NLI task.

[6] We analyse the impact of fine-tuning on perturbed examples in §6.8.

the label is expected to suffer on the label-flipped cases. In case of label-preserving perturbations of similar kind, structural correlations between the hypothesis and the label are retained leading to minimal drop in model performance.

The results of the hypothesis-only model on the *Train* set may appear slightly surprising at first. However, given that the model was trained on this dataset, it seems reasonable to assume that the model has 'overfit' to the training data. Therefore, the model is expected to be vulnerable even to slight label-preserving modifications to the examples it was trained on, leading to the huge drop of 26%. In the same setting, for the α_1 set the performance drop is lesser, namely about 3%.

Taken together, we can conclude from these results that the model ignores the information in the hypotheses, (thereby perhaps also the aligned facts in the premise), and instead relies on irrelevant structural patterns in the hypotheses.

6.6 Probing Evidence Selection

Predictions of an NLI model should primarily be based on the evidence in the premise, that is, on the facts relevant to the hypothesis. For a tabular premise, rows containing the evidence necessary to infer the associated hypothesis are called relevant rows. Short-circuiting the evidence in relevant rows for inference using annotation artifacts as suggested in §6.5 or other spurious artifacts in irrelevant rows of the table is expected to lead to poor generalization over unseen data.

6.6.1 Assumptions: Facts versus Common Sense

Primary Assumption: When a row is deleted or updated in a table, the truth value of the corresponding fact(s) in that row may change from true/false to undetermined. This assumption holds under the condition that each row in the table represents an independent fact that is mutually exclusive of other rows. This assumption is generally applicable to entity-centric tables, but it may not always hold true. Facts in a table are not necessarily common sense, but they represent knowledge that can be obtained from a reliable source. The distinction between common sense and knowledge is not always clear, and there may be cases where the line between the two needs to be crossed.

When Assumptions Break a.k.a. the Exceptional Cases: The assumption of indepen-

dent and mutually exclusive rows may break in certain cases, such as when dealing with exceptions or edge cases that are not covered in common sense. In such cases, it is important to be careful and to consider the nature of the facts being represented in the table. The definition of what constitutes common knowledge versus non-common knowledge can be a topic of debate, and the existence of a clear boundary may be uncertain. For examples Platypus is a rare mammal which can lay eggs, thus removing the knowledge of mammal from his table, can create confusion.

To better understand the model's ability to select evidence in the premise, we use two kinds of controlled edits: (a) automatic edits without any information about relevant rows, and, (b) semi-automatic edits using knowledge of relevant rows via manual annotation. The rest of the section goes over both scenarios in detail. All experiments in this section use the full model that is trained on both premises and their associated hypotheses.

6.6.2 Automatic Probing

We define four kinds of table modifications that are agnostic to the relevance of rows to a hypothesis: (a) *row deletion*, (b) *row insertion*, (c) *row-value update*, i.e., changing existing information, and (d) *row permutation*, i.e., reordering rows. Each modification allows certain desired (valid) changes to model predictions.[7] We examine below the case of row deletion in detail and refer the reader to the Appendix for the others.

6.6.2.1 Row deletion

Row deletion should lead to the following desired effects: (a) If the deleted row is relevant to the hypothesis (e.g., *Length* for H1), the model prediction should change to NEUTRAL. (b) If the deleted row is irrelevant (e.g., *Producer* for H1), the model should retain its original prediction. NEUTRAL predictions should remain unaffected by row deletion.

Results and analysis: We studied the impact of *row deletion* on the α_1, α_2 and α_3 test sets. Figure 6.2 shows aggregate changes to labels after row deletions as a directed labeled graph. The nodes in this graph represent the three labels in INFOTABS, and the edges

[7] In performing these modifications, we made sure that the modified table does not become inconsistent or self-contradicting.

denote transitions after *row deletion*. The source and end nodes of an edge represent predictions before and after the modification.

We see that the model makes invalid transitions in all three datasets. Table 6.6 summarizes the invalid transitions by aggregating them over the label originally predicted by the model. The percentage of invalid transitions is higher for ENTAIL predictions than for CONTRADICT and NEUTRAL. After row deletion, many ENTAIL examples are incorrectly transitioning to CONTRADICT rather than to NEUTRAL. The opposite trend is observed for the CONTRADICT predictions.

6.6.2.2 Row insertion

When we insert new information that does not contradict an existing table,[8] original predictions should be retained in almost all cases. Very rarely, NEUTRAL labels may change to ENTAIL or CONTRADICT. For example, adding the *Singles* row below to our running example table doesn't change labels for any hypothesis except the H4 label (see Table 6.1) changing to CONTRADICT with the additional information.

Singles | The Logical Song; Breakfast in America; Goodbye Stranger; Take the Long Way Home

Results and Analysis: Figure 6.3 shows the possible label changes after new row insertion as a directed labeled graph, and the results are summarized in Table 6.7. Note that all transitions from NEUTRAL are valid upon row insertion, although not all may be accurate.

6.6.2.3 Row update

In case of row update, we only change a portion of a row value. Whole row value substitutions are examined separately as composite operations of deletion followed by insertion. Unlike a whole row update, changing only a portion of a row is non-trivial. We must ensure that the updated value is appropriate for the key in question and also avoid self-contradictions. To satisfy these constraints, we update a row with a value from a random table with the same key and only update values in multi-valued rows. A row update operation may have an effect on all labels. Though feasible, we consider the

[8] To ensure that the information added is not contradictory to existing rows, we only add rows with new keys instead of changing values for the existing keys.

transitions from CONTRADICT to ENTAIL to be prohibited. Unlike ENTAIL to CONTRADICT transitions, these transitions would be extremely rare as values are updated randomly, regardless of their semantics. For example, if we substitute *pop* in the multi-valued key *Genre* in our running example with another genre, the hypothesis H1 is likely to change to CONTRADICT.

Results and Analysis: The model mostly retains its predictions on row-value update operations. Since we are updating a single value from a multi-valued key, the changes to the table are minimal and may not be perceived by the model. As a result, we should expect row updates to have lower impact on model predictions. This appears to be the case, as evidenced by the results in Figure 6.4, which show that the labels do not change drastically after update. The results in Figure 6.4 are summarized in Table 6.8.

6.6.2.4 Row permutation

By design of the premises, the order of their rows should have no effect on hypotheses labels. In other words, the labels should be invariant to row permutation.

Results and Analysis: However, from Figure 6.5, it is evident that even a simple shuffling of rows, where no information has been tampered with, can have a notable effect on performance. This shows that the model is relying on row positions incorrectly, while the semantics of a table is order invariant. We summarize the combined invalid transitions from Figure 6.5 in Table 6.9. This suggest some form of position bias in the model.

6.6.2.5 Composition of perturbation operations

In addition to probing individual operations, we can also study their compositions. For example, we could delete a row, and insert a different row, and so on. The composition of these operations have interesting properties with respect to the allowed transitions. For example, when an operation is composed with itself (e.g. two deletions), the set of valid label changes is the same as for the operation. A particularly interesting composition is deletion followed by an insertion, since this can be viewed as a row update.

Results and Analysis: In Figure 6.6, we show the transition graph for the composition operation of row deletion followed by insertion and the summary of the possible transitions is presented in Table 6.10.

6.6.3 Manual Probing

Row modification for automatic probing in §6.6.2 is agnostic to the relevance of the row to a given hypothesis. Since only a few rows (one or two) are relevant to the hypothesis, the probing skew towards hypothesis-unrelated rows weakens the investigations into the evidence-grounding capability of the model. Knowing the relevance of rows allows for the creation of stronger probes. For example, if a relevant row is deleted, the ENTAIL and CONTRADICT predictions should change to NEUTRAL. (Recall that after deleting an irrelevant row the model should retain its original label.)

Probing by altering or deleting relevant rows requires human annotation of relevant rows for each table-hypothesis pair. We used MTurk to annotate the relevance of rows in the development and the test sets, with turkers identifying the relevant rows for each table-hypothesis pair.

6.6.3.1 Inter-annotator agreement

We employed majority voting to derive ground truth labels from multiple annotations for each row. The inter-annotator agreement macro F1 score for each of the four datasets is over 90% and the average Fleiss' kappa is 78 (std: 0.22). This suggests good inter-annotator agreement. In 82.4% of cases, at least 3 out of 5 annotators marked the same relevant rows.

6.6.3.2 Results and analysis

We examined the response of the model when relevant rows are deleted. Figure 6.7 shows the label transitions. The fact that even after the deletion of relevant rows, ENTAIL and CONTRADICT predictions don't change to NEUTRAL a large percentage of times (mostly the original label remains unchanged and at other times, it changes incorrectly), indicates that the model is likely utilizing spurious statistical patterns in the data for making the prediction.

We summarize the combined invalid transitions for each label in Table 6.11. We see that the percentage of invalid transitions is considerably higher compared to random row deletion in Figure 6.2.[9] The large percentage of invalid transitions in the ENTAIL and CONTRADICT cases indicates a rather high utilization of spurious statistical patterns by

[9]Note that the dashed black lines from Figure 6.2 are now red in Figure 6.7, indicating invalid transitions.

the model to arrive at its answers.

Irrelevant Row Deletion: Ideally, deletion of an irrelevant row should have no effect on a hypothesis label. The results in Figure 6.8 and in Table 6.12 show that even irrelevant rows have an effect on model predictions. This further illustrates that the seemingly accurate model predictions are not appropriately grounded on evidence.

6.6.4 Human versus Model Evidence Selection

We further analyze the model's capability for selecting relevant evidence by comparing it with human annotators. All rows that alter the model predictions during automatic row deletion are considered as *model relevant rows* and are compared to the human-annotated relevant rows. We only consider the subset of 4600 (from 7200 annotated dev/test sets pairs) hypothesis-table pairs with ENTAIL and CONTRADICT labels, where deleting a relevant row should change the prediction to NEUTRAL.[10]

6.6.4.1 Results and analysis

On the human-annotated relevant rows, the model has an average precision of 41.0% and a recall of 40.9%. Further analysis reveals that the model (a) uses all relevant rows in 27% cases, (b) uses incorrect or no rows as evidence in 52% of occurrences, and (c) is only partially accurate in identifying relevant rows in the remaining 21% of examples. Upon further analysing the cases in (b), we observed that the model actually ignores premises completely in 88% (of 52%) of cases. This accounts for 46% (absolute) of all occurrences. In comparison, in the human-annotated data, such cases only amount to $< 2\%$.

Although, the model's predictions are 70% correct in the 4,600 examples, only 21% can be attributed to using all relevant evidence. The correct label in 37% of the 4,600 examples is from irrelevant rows, with the remaining 12% of correct predictions use some, but not all, relevant rows. We can conclude from the findings in this section that the model does not seem to need all the relevant evidence to arrive at its predictions, raising questions about trust in its predictions.

---

[10]We did not include the 2400 NEUTRAL examples pairs and the ambiguous 200 ENTAIL or CONTRADICT examples that had no relevant rows as per the consensus annotation.

6.7 Probing with Counterfactual Examples

Since INFOTABS is a dataset of facts based on Wikipedia, pre-trained language models such as RoBERTa, trained on Wikipedia and other publicly available text, may have already encountered information in INFOTABS during pre-training. As a result, NLI models built on top of RoBERTa$_L$ can learn to infer a hypothesis using the knowledge of the pre-trained language model. More specifically, the model may be relying on "*confirmation bias*", in which it selects evidence/patterns from both premise and hypothesis that matches its prior knowledge. While world knowledge is necessary for table NLI [182], models should still treat the premise as the primary evidence.

Counterfactual examples can help test whether the model is grounding its inference on the evidence provided in the tabular premise. In such examples, the tabular premise is modified such that the content does not reflect the real world. In this study, we limit ourselves to modifying only the ENTAIL and CONTRADICT examples. We omit the NEUTRAL cases because the majority of them in INFOTABS involve out-of-table information; producing counterfactuals for them is much harder and involves the laborious creation of new rows with the right information.

The task of creating counterfactual tables presents two challenges. First, the modified tables should not be self-contradictory. Second, we need to determine the labels of the associated hypotheses after the table is modified. We employ a simple approach to generate counterfactuals that addresses both challenges. We use the evidence selection data (§6.6.3) to gather all premise-hypothesis pairs that share relevant keys such as "Born", "Occupation" etc. Counterfactual tables are generated by swapping the values of relevant keys from one table to another.[11]

Figure 6.9 shows an example. We create counterfactuals from the premises in Table 6.1 and Figure 6.1 by swapping their **Length** rows. We also swap the hypotheses (H1 and H5) aligned to the **Length** rows in both premises by replacing the title expression **Bridesmaids** in H5 with **Breakfast in America** and *vice versa*. The simple procedure ensures that the hypotheses labels are left unchanged in the process, resulting in high-quality data.

In addition, we also generated counterfactuals by swapping the table title and associ-

[11]There may still be a few cases of self-contradiction, but we expect that such invalid cases would not exist in the rows that are relevant to the hypothesis.

ated expressions in the hypotheses with the title of another table, resulting in a counterfactual table-hypothesis pair, as in the row swapping strategy. Figure 6.10 shows an example created from the premises in Table 6.1 and Figure 6.1 by swapping the title rows **Breakfast in America** and **Bridesmaids**. The title expression in all hypotheses in Table 6.1 are also replaced by **Bridesmaids**. This strategy also preserves the hypothesis label similar to row swapping.

The above approaches are *Label Preserving* as they do not alter the entailment labels. Counterfactual pairs with flipped labels are important for filtering out the contribution of artifacts or other spurious correlations that originate from a hypothesis (see §6.5). So, in addition, we also created counterfactual table-hypothesis pairs where the original labels are flipped. These counterfactual cases are, however, non-trivial to generate automatically, and are therefore created manually. To create the *Label-Flipped* counterfactual data, three annotators manually modified tables from the *Train* and α_1 datasets corresponding to ENTAIL and CONTRADICT labels, producing 885 counterfactual examples from the *Train* set and 942 from the α_1 set. The annotators cross-checked the labels to determine annotation accuracy, which was 88.45% for the *Train* set and 86.57% for the α_1 set.

6.7.1 Results and Analysis

We tested both hypothesis-only and full (Prem+Hypo) models on the counterfactual examples created above, without fine-tuning on a subset of these examples. The results are presented in Table 6.13 where each cell represents average accuracy and standard deviation (subscript) over 100 sets of 80% randomly sampled counterfactual examples. We see that the (Prem+Hypo) model is not robust to counterfactual perturbations. On the label-flipped counterfactuals, the performance drops down to close to a random prediction (48.70% for *Train* and 44.01% for α_1). The performance on the label-preserved counterfactuals is relatively better which leads us to conjecture that the model largely exploits artifacts in hypotheses.

Due to over-fitting, the *Train* set has a larger drop of 15.85%, compared to only 2.70% on the α_1 set on the label-preserved examples. Moreover, the drop in performance for both *Prem+Hypo* and *Hypo-Only* models is comparable to their performance drop on the original table-hypothesis pairs. This shows that, regardless of whether the relevant information in

the premise is accurate, both models rely substantially on hypothesis artifacts. On the *Label-Flipped* counterfactuals, the large drop in accuracy could be due to both ambiguous hypothesis artifacts or counterfactual information.

To disentangle these two factors, we can take advantage of the fact that the counter-factual examples are constructed from, and hence paired with, the original examples. This allows us to examine pairs of examples where the full model makes an incorrect prediction on one, but not the other. Especially of interest are the cases where the full model makes a correct prediction on the original example, but not on the corresponding counterfactual example.

Table 6.14 shows the results of this analysis. Each row represents a condition corresponding to whether the full and the hypothesis-only models are correct on the original example. The two cases of interest, described above, correspond to the second and fourth rows of the table. The second row shows the case where the full model is correct on the original example (and not on the counter-factual example), but the hypothesis-only model is not. Since we can discount the impact of hypothesis bias in these examples, the error in the counter-factual version could be attributed to reliance on pre-trained knowledge. Unsurprisingly, there are no such examples in the training set. In the α_1 set, we see a substantial fraction of counterfactual examples (11.79%) belong to this category. The last row considers the case where the hypothesis-only model is correct. We see that this accounts for a larger fraction of the counterfactual errors, both in the training and the α_1 sets. Among these examples, *despite* the (albeit unfortunate) fact that the hypothesis alone can support a correct prediction, the model's reliance on its pre-trained knowledge leads to errors in the counterfactual cases.

The results, taken in aggregate, suggest that the model produces predictions based on hypothesis artifacts and pre-trained knowledge rather than the evidence presented to it, thus impacting its robustness and generalization.

6.8 Inoculation by Fine-Tuning

Our probing experiments demonstrate that the models, trained on the INFOTABS training set, failed along all three dimensions that we investigated. This leads us to the following question: *Can additional fine-tuning with perturbed examples help?*

[155] point out that poor performance on challenging datasets can be ascribed to either a weakness in the model, a lack of diversity in the dataset used for training or information leakage in the form of artifacts.[12] They suggest that models can be further fine-tuned on a few challenging examples to determine the possible source of degradation. Inoculation can lead to one of three outcomes: (a) **Outcome 1:** The performance gap between the challenge and the original test sets reduces, possibly due to addition of diverse examples, (b) **Outcome 2:** Performance on both the test sets remains unchanged, possibly because of the model's inability to adapt to the new phenomena or the changed data distribution, or, (c) **Outcome 3:** Performance degrades on the test set, but improves on the challenge set, suggesting that adding new examples introduces ambiguity or contradictions.

We conducted two sets of inoculation experiments to help categorize performance degradation of our models into one of these three categories. For each experiment described below, we generated additional inoculation datasets with 100, 200 and 300 examples to inoculate the original task-specific RoBERTa$_L$ models trained on both premises and hypotheses. As in the original inoculation work, we created these adversarial datasets by sub-sampling inclusively, i.e., the smaller datasets are subsets of the larger ones. Following the training protocol in [155], we tried learning rates of 10^{-4}, 5×10^{-5} and 10^{-5}. We performed inoculation for a maximum of 30 epochs with early stopping based on the development set accuracy. We found that with the first two learning rates, the model does not converge, and underperforms on the development set. The model performance was best with the learning rate of 10^{-5}, which we used throughout the inoculation experiments. The standard deviation over 100 sample splits for all experiments was ≤ 0.91.

6.8.1 Annotation Artifacts

Table 6.15 shows the results on the hypothesis-perturbed examples (from §6.5), and Table 6.16 shows the performance of the inoculated models on the original INFOTABS test sets. We see that fine-tuning on the hypothesis-perturbed examples decreases performance on the original α_1, α_2 and α_3 test sets, but performance improves on the more difficult label-flipped examples of the hypothesis-perturbed test set.

[12] Model weakness is the inherent inability of a model (or a model family) to handle certain linguistic phenomena.

6.8.2 Counterfactual Examples

Tables 6.17 and 6.18 show the performance of models inoculated on the original INFOTABS test sets and the counterfactual examples from §6.7 respectively. Once again, we see that fine-tuning on counterfactual examples improves performance on the adversarial counterfactual examples test set, at the cost of performance on the original test sets.

6.8.3 Analysis

We see that both experiments above belong to Outcome 3, where the performance improves on the challenge set, but degrades on the test set(s). The change in the distribution of inputs hurts the model: we conjecture that this may be because the RoBERTa$_L$ model exploits data artifacts in the original dataset but fails to do so for the challenge dataset and vice versa.

We expect our model to handle both original and challenge datasets, at least after fine-tuning (i.e., it should belong to Outcome 1). Its failure points to the need for better models or training regimes.

6.9 Discussion and Related Work

6.9.1 What Did We Learn?

Firstly, through systematic probing, we have shown that despite good performance on the evaluation sets, the model for tabular NLI fails at reasoning. From the analysis of hypothesis perturbations (§6.5), we show that the model heavily relies on correlations between a hypothesis' sentence structure and its label. Models should be systematically evaluated on adversarial sets like α_2 for robustness and sensitivity. This observation is concordant with multiple studies that probe deep learning models on adversarial examples in a variety of tasks such as question answering, sentiment analysis, document classification, natural language inference, etc. [79, 139, 221, 222, 253].

Secondly, the model does not look at correct evidence required for reasoning, as is evident from the evidence-selection probing (§6.6). Rather, it leverages spurious patterns and statistical correlations to make predictions. A recent study by [139] on question-answering shows that models indeed leverage spurious patterns to answer a large fraction (60-70%) of questions.

Thirdly, from counterfactual probes (§6.7), we found that the model relies on knowledge of pre-trained language models than on tabular evidence as the primary source of knowledge for making predictions. This is in addition to the spurious patterns or hypothesis artifacts leveraged by the model. Similar observations are made by [38, 72, 99, 108, 119, 154, 259, 272, 299] for unstructured text.

Finally, from the inoculation study (§6.8), we found that fine-tuning on challenge sets improves model performance on challenge sets but degrades on the original α_1, α_2, and α_3 test sets. That is, changes in the data distribution during training have a negative impact on model performance. This adds weight to the argument that the model relies excessively on data artifacts.

6.10 Conclusion

This chapter presented a targeted probing study to highlight the limitations of tabular inference models using a case study on a tabular NLI task on INFOTABS. Our findings show that despite good performance on standard splits, a RoBERTa-based tabular NLI model, fine-tuned on the existing pre-trained language model, fails to select the correct evidence, makes incorrect predictions on adversarial hypotheses, and is not grounded in provided evidence–counterfactual or otherwise. We expect that insights from the study can help in designing rationale selection techniques based on structural constraints for tabular inference and other tasks. While inoculation experiments showed partial success, diverse data augmentation may help mitigate challenges. However, annotation of such data can be expensive. It may also be possible to train models to satisfy domain-based constraints [142] to improve model robustness. Finally, probing techniques described here may be adapted to other NLP tasks involving tables such as tabular question answering and tabular text generation.

	Bridesmaids
Length	125 minutes

H5: Bridesmaids has a running time of over 3 hrs.

Figure 6.1: Hypothesis H5 contradicts the premise.

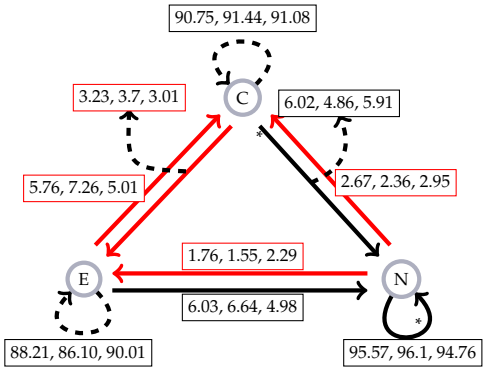


Figure 6.2: Changes in model prediction after row deletion. Red lines represent invalid transitions. Dashed and solid lines represent valid transitions for irrelevant and relevant row deletion respectively. * represents valid transitions with either row deletions. The edge labels represents the percentage of transitions for α_1, α_2 and α_3 set in order.

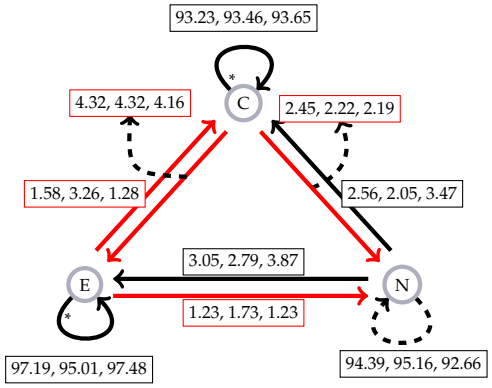


Figure 6.3: Changes in model predictions after new row insertion. (Notation similar to Figure 6.2)

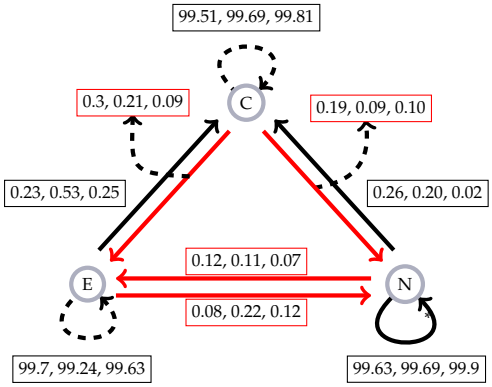


Figure 6.4: Changes in model predictions after row value update. (Notation similar to Figure 6.2)

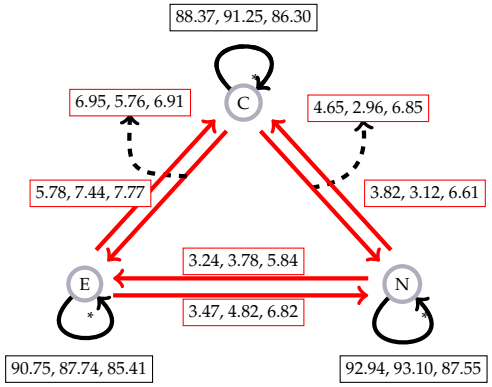


Figure 6.5: Changes in model predictions after shuffling of table rows. (Notation similar to Figure 6.2.)

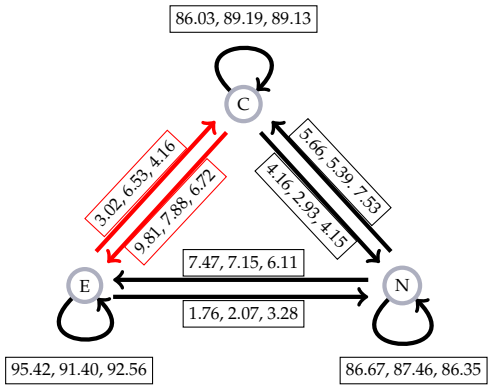


Figure 6.6: Changes in model predictions after deletion followed by an insert operation. (Notation similar to Figure 6.2.)

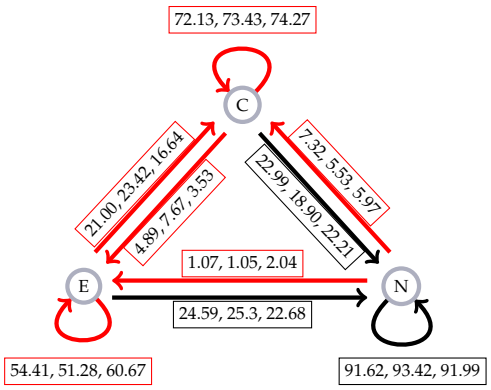


Figure 6.7: Changes in model prediction after deletion of **relevant rows**. Red lines represent invalid transitions while solid lines represent valid transitions. The edge labels represent transitions for α_1, α_2 and α_3 set in order.

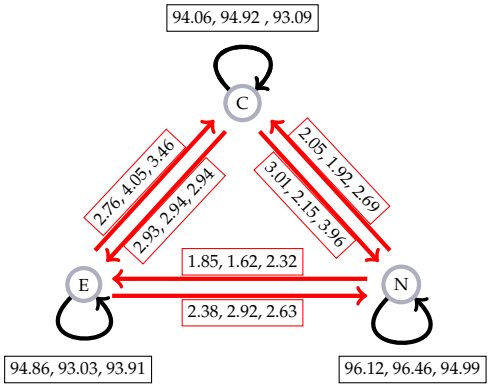


Figure 6.8: Change in model predictions after deletion of an **irrelevant row**. (Notation similar to Figure 6.2.)

Breakfast in America	
Released	29 March 1979
Recorded	May–December 1978
Studio	The Village Recorder (Studio B) in Los Angeles
Genre	pop ; art rock ; soft rock
Length	125 minutes
Label	A&M
Producer	Peter Henderson, Supertramp

$\widehat{H5}$: Breakfast in America has a running time of over 3 hrs.

Figure 6.9: Counterfactual table-hypothesis pair created from Table 6.1 and Figure 6.1. Only the values of 'Length' rows are swapped, rest of the rows from Table 6.1 are copied as such.

Bridesmaids	
Released	29 March 1979
Recorded	May–December 1978
Studio	The Village Recorder (Studio B) in Los Angeles
Genre	pop ; art rock ; soft rock
Length	46:06
Label	A&M
Producer	Peter Henderson, Supertramp

$\widehat{H1}$: Bridesmaids is a pop album with a length of 46 minutes.

$\widehat{H2}$: Bridesmaids was released at the end of 1979.

$\widehat{H3}$: Most of Bridesmaids was recorded in the last month of 1978.

$\widehat{H4}$: Bridesmaids has 6 tracks.

Figure 6.10: A counterfactual tabular premise and the associated hypotheses created from Table 6.1. The hypotheses $\widehat{H1}$ is entailed by the premise, $\widehat{H2}$ contradicts it, and $\widehat{H3}$ and $\widehat{H4}$ are neutral.

Table 6.1: A tabular premise example. The hypotheses H1 is entailed by it, H2 contradicts it, and H3, H4 are neutral i.e. neither entailed nor contradictory.

Breakfast in America	
Released	29 March 1979
Recorded	May–December 1978
Studio	The Village Recorder (Studio B) in Los Angeles
Genre	pop ; art rock ; soft rock
Length	46:06
Label	A&M
Producer	Peter Henderson, Supertramp

H1: Breakfast in America is a pop album with a length of 46 minutes.

H2: Breakfast in America was released at the end of 1979.

H3: Most of Breakfast in America was recorded in the last month of 1978.

H4: Breakfast in America has 6 tracks.

Table 6.2: Results of the *Table as a Paragraph* strategy on INFOTABS subsets with RoBERTa$_L$ model, hypothesis-only baseline and majority human agreement. The first three rows are reproduced from [84]. The last row represents the average performances (and standard deviations as subscripts) using models obtained via five-fold cross validation (5xCV).

Model	dev	α_1	α_2	α_3
Human	**79.78**	**84.04**	**83.88**	**79.33**
Hypothesis Only	60.51	60.48	48.26	48.89
RoBERTa$_L$	75.55	74.88	65.55	64.94
5xCV	$73.59_{(2.3)}$	$72.41_{(1.4)}$	$63.02_{(1.9)}$	$61.82_{(1.4)}$

Table 6.3: Monotonicity properties of prepositions.

Preposition	**Upward Monotonicity**	**Downward Monotonicity**
over	CONTRADICT	ENTAIL
under	ENTAIL	CONTRADICT
more than	CONTRADICT	ENTAIL
less than	ENTAIL	CONTRADICT
before	ENTAIL	CONTRADICT
after	CONTRADICT	ENTAIL

Table 6.4: Example hypothesis perturbations for the running example from Table 6.1. The *red italicized* text represent changes. Superscripts E/C represent gold ENTAIL and CONTRADICT labels, while subscripts E/C represent new labels.

Type of Modification	**Perturbed Hypothesis**
Nominal Modifier	H1^E_E: Breakfast in America *which was produced by Pert Handerson* is a pop album of 46 minutes length.
Temporal Expression	H1^E_C: Breakfast in America is a pop album with a length of *56* minutes.
Negation	H2^C_E: Breakfast in America was *not* released towards the end of 1979.
Temporal Expression	H2^C_C: Breakfast in America was released towards the end of *1989*.

Table 6.5: Results of the Hypothesis-only model and Prem-Hypo RoBERTa$_L$ model (trained on hypothesis and premise together) on the gold and perturbed examples from the *Train* and α_1 sets. Main numbers are the mean and subscript$_{(.)}$ is corresponding std. calculated with 80% data random splits over 100 times.

Model	**Original**	**Label Preserved**	**Label Flipped**
		Train **Set (w/o NEUTRAL)**	
Prem+Hypo	$99.44_{(0.06)}$	$92.98_{(0.20)}$	$53.92_{(0.28)}$
Hypo-Only	$96.39_{(0.13)}$	$70.23_{(0.35)}$	$19.23_{(0.27)}$
		α_1 **Set (w/o NEUTRAL)**	
Prem+Hypo	$68.94_{(0.76)}$	$69.56_{(0.77)}$	$51.48_{(0.86)}$
Hypo-Only	$63.52_{(0.75)}$	$60.27_{(0.85)}$	$31.02_{(0.63)}$

Table 6.6: Percentage of invalid transitions after row deletion.

Dataset	α_1	α_2	α_3	Average
ENTAIL	5.76	7.26	5.01	6.01
NEUTRAL	4.43	3.91	5.24	4.53
CONTRADICT	3.23	3.70	3.01	3.31
Average	4.47	4.96	4.42	-

Table 6.7: Percentage of invalid transitions after new row insertion. For an ideal model, all these numbers should be zero.

Dataset	α_1	α_2	α_3	Average
ENTAIL	2.81	4.99	2.51	3.44
NEUTRAL	0	0	0	0
CONTRADICT	6.77	6.54	6.35	6.55
Average	3.19	3.84	2.95	-

Table 6.8: Percentage of invalid transitions after row value update. For an ideal model, all these numbers should be zero.

Dataset	α_1	α_2	α_3	Average
ENTAIL	0.08	0.22	0.12	0.14
NEUTRAL	0.12	0.11	0.09	0.11
CONTRADICT	0.49	0.30	0.19	0.33
Average	0.23	0.21	0.13	-

Table 6.9: Percentage of invalid transitions after row permutations. For an ideal model, all these numbers should be zero.

Dataset	α_1	α_2	α_3	Average
ENTAIL	9.25	12.2	14.6	12.02
NEUTRAL	7.1	6.8	12.5	8.79
CONTRADICT	11.6	8.76	13.7	11.36
Average	9.34	9.26	13.6	-

Table 6.10: Percentage of invalid transitions after deletion followed by an insertion operation. For an ideal model, all these numbers should be zero.

Datasets	α_1	α_2	α_3	Average
ENTAIL	3.02	6.53	4.16	4.57
NEUTRAL	0.00	0.00	0.00	0.00
CONTRADICT	9.81	7.88	6.71	8.13
Average	4.28	4.80	3.63	-

Table 6.11: Percentage of invalid transitions following deletion of relevant rows. For an ideal model, all these numbers should be zero.

Dataset	α_1	α_2	α_3	Average
ENTAIL	75.41	74.70	77.31	75.80
NEUTRAL	8.39	6.58	8.01	7.66
CONTRADICT	77.02	81.10	77.80	78.64
Average	53.60	54.14	54.35	

Table 6.12: Percentage of invalid transitions after deletion of irrelevant rows. For an ideal model, all these numbers should be zero.

Datasets	α_1	α_2	α_3	Average
ENTAIL	5.14	6.97	6.09	6.07
NEUTRAL	3.9	3.54	5.01	4.15
CONTRADICT	5.94	5.09	6.91	5.98
Average	4.99	5.2	6.01	-

Table 6.13: Results of the Hypothesis-only and Prem-Hypo RoBERTa$_L$ models (trained on hypothesis and premise together) on the gold and counterfactual examples from the *Train* and α_1 sets. Each entry denotes the $mean_{(stdev)}$ calculated with 80% data random splits over 100 times.

Model	Original	Label Preserved	Label Flipped
Train **Set (w/o NEUTRAL)**			
Prem+Hypo	$94.38_{(0.39)}$	$78.53_{(0.65)}$	$48.70_{(0.72)}$
Hypo-Only	$99.94_{(0.06)}$	$82.23_{(0.65)}$	$00.06_{(0.01)}$
α_1 **Set (w/o NEUTRAL)**			
Prem+Hypo	$71.99_{(0.69)}$	$69.65_{(0.78)}$	$44.01_{(0.72)}$
Hypo-Only	$60.89_{(0.76)}$	$58.19_{(0.91)}$	$27.68_{(0.65)}$

Table 6.14: Comparison of results of the full and hypothesis-only models on the original and counterfactual examples. O-THP and C-THP represent original and counterfactual table-hypothesis pairs, O-Hypo represents hypotheses from the original data, ✓ represents correct predictions and ✗ represents incorrect predictions.

Prem+Hypo		Hypo-Only	Dataset	
C-THP	O-THP	O-Hypo	*Train*	α_1
✓	✗	✗	0.00	11.43
✗	✓	✗	0.00	11.79
✓	✗	✓	3.57	6.48
✗	✓	✓	49.36	33.12

Table 6.15: Performance of the inoculated models on the hypothesis perturbed INFOTABS sets. Variance across 100 sample splits was ≤ 0.91.

#Samples	Original	Label Preserved	Label Flipped
Train **Set (w/o NEUTRAL)**			
0 (w/o Ino)	**99.44**	92.98	53.92
100	97.24	95.58	**79.25**
200	97.24	**95.65**	78.75
300	97.24	95.64	78.74
α_1 **Set (w/o NEUTRAL)**			
0 (w/o Ino)	**68.94**	**69.56**	51.48
100	68.05	65.67	**57.91**
200	68.37	66.29	57.49
300	68.36	66.29	57.49

Table 6.16: Performance of the inoculated models on the original INFOTABS test sets. Variance across 100 sample splits was ≤ 0.91.

#Samples	α_1	α_2	α_3
0 (w/o Ino)	**74.88**	**65.55**	**64.94**
100	67.44	62.17	58.51
200	67.34	61.88	58.61
300	67.24	61.84	58.62

Table 6.17: Performance after inoculation by fine-tuning on the original INFOTABS test sets.

#Samples	α_1	α_2	α_3
0 (w/o Ino)	**74.88**	**65.55**	**64.94**
100	69.72	63.88	59.66
200	69.88	63.78	58.89
300	67.34	62.23	57.58

Table 6.18: Performance after inoculation fine-tuning on the INFOTABS counterfactual example sets.

#Samples	Original	Label Preserved	Label Flipped
Train **Set (w/o NEUTRAL)**			
0 (w/o Ino)	**94.38**	78.53	48.70
100	91.82	84.61	57.62
200	92.46	**84.92**	**59.43**
300	91.08	83.54	63.58
α_1 **Set (w/o NEUTRAL)**			
0 (w/o Ino)	**71.99**	69.65	44.01
100	66.05	75.03	50.40
200	65.86	**75.03**	50.57
300	65.59	74.23	**52.09**

CHAPTER 7

TRUSTWORTHY TABULAR REASONING

Adapted from V. Gupta, S. Zhang, A. Vempala, Y. He, T. Choji, and V. Srikumar, *Right for the right reason: Evidence extraction for trustworthy tabular reasoning*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, pp. 3268–3283.

In Chapter 6, we demonstrates that tabular reasoning models fail to reason properly on the semi-structured inputs in many cases. For example, they can ignore relevant rows, and (a) focus on the irrelevant rows [182], (b) use only the hypothesis sentence [87, 202], or (c) knowledge acquired during pre-training [83, 104]. In essence, they use spurious correlations between irrelevant rows, the hypothesis, and the inference label to predict labels.

This chapter argues that existing NLI systems optimized solely for label prediction cannot be trusted. It is not sufficient for a model to be merely *Right* but also *Right for the Right Reasons*. In particular, at least identifying the relevant elements of inputs as the *'Right Reasons'* is essential for trustworthy reasoning[1]. We address this issue by introducing the task of *Trustworthy Tabular Inference*, where the goal is to extract relevant rows as evidence and predict inference labels.

To illustrate this task, consider an example from the INFOTABS dataset in Table 7.1, which shows a premise table and three hypotheses. The figure also marks the rows needed to make decisions about each hypothesis, and also indicates the relevant tokens for each hypothesis. For trustworthy tabular reasoning, in addition to predicting the label ENTAIL for *H1*, CONTRADICT for *H2* and NEUTRAL for *H3*, the model should also identify the evidence rows—namely, the rows *Producer* and *Length* for hypothesis *H1*, *Recorded* for hypothesis *H2*, *Released* and *Recorded* for hypothesis *H3*.

[1]We argue that a reasoning system can be deemed trustworthy only if it exposes how its decisions are made, thus admitting verification of the reasons for its decisions.

As a first step, we propose a two-stage sequential prediction approach for the task, comprising of an evidence extraction stage, followed by an inference stage. In the evidence extraction stage, the model extracts the necessary information needed for the second stage. In the inference stage, the NLI model uses only the extracted evidence as the premise for the label prediction task.

We explore several unsupervised evidence extraction approaches for INFOTABS. Our best unsupervised evidence extraction method outperforms a previously developed baseline by 4.3%, 2.5% and 5.4% absolute score on the three test sets. For supervised evidence extraction, we annotate the INFOTABS training set (17K table-hypothesis pairs with 1740 unique tables) with relevant rows following the methodology of [83], and then train a RoBERTa$_{\text{LARGE}}$ classifier. The supervised model improves the evidence extraction performance by 8.7%, 10.8%, and 4.2% absolute scores on the three test sets over the unsupervised approach. Finally, for the full inference task, we demonstrate that our two-stage approach with best extraction, outperforms the earlier baseline by 1.6%, 3.8%, and 4.2% on the three test sets. This work is published at ACL 2022 as [85].

7.1 Contributions

We make the following contributions in this chapter [2]:

1. We introduce the problem of trustworthy tabular reasoning and study a two-stage prediction approach that first extracts evidence and then predicts the NLI label.
2. We investigate a variety of unsupervised evidence extraction techniques. Our unsupervised approach for evidence extraction outperforms the previous methods.
3. We enrich the INFOTABS training set with evidence rows, and develop a supervised extractor that has near-human performance.
4. We demonstrate that our two-stage technique with best extraction outperforms all the prior benchmarks on the downstream NLI task.

7.2 Background

7.2.1 Interpretability and Explainability

Model interpretability can either be through explanations or by identifying the evidence for the predictions [52, 63, 105, 192, 234, 279]. Additionally, NLI models [78, 103, 155,

[2]The updated dataset, along with associated code, is available at https://tabevidence.github.io/.

167, 179, 184, 218, 219, 220, 307] must be subjected to numerous test sets with adversarial settings. These settings can focus on various aspects of reasoning, such as perturbed premises for evidence selection [83], zero-shot transferability (α_3), counterfactual premises [104], and contrasting hypotheses α_2. Recently, [130] introduced Natural-language Inference over Label-specific Explanations (NILE), an NLI approach for generating labels and accompanying faithful explanations using auto-generated label-specific natural language explanations. Our work focuses on the extraction of label-independent evidence for correct inference, rather than on the generation of abstractive explanations for a given label.

7.2.2 Tabular Shared Tasks

The SemEval'21 Task 9 [226] and FEVEROUS'21 shared task [5] are conceptually close to this work.

The SemEval task focuses on statement verification and evidence extraction using relational tables from scientific articles. In this work, we focus on item evidence extraction for non-scientific Wikipedia Infobox entity tables, proposed a two-stage sequential approach, and used the INFOTABS dataset which has complex reasoning and multiple adversarial tests for robust evaluation.

The FEVEROUS'21 shared task focuses on verifying information using unstructured and structured evidence from open-domain Wikipedia. Our approach concerns evidence extraction from a single table rather than open-domain document, table or paragraph retrieval. Furthermore, we are only concerned with entity tables rather than relational tables or unstructured text, while the FEVEROUS data has relational tables, unstructured text, and fewer entity tables.

7.3 Task Formulation

We begin by introducing the task and the datasets we use.

7.3.1 Tabular Inference

Tabular inference is a reasoning task that, like conventional NLI [17, 45, 280], asks whether a natural language *hypothesis* can be inferred from a tabular *premise*. Concretely, given a premise table T with m rows $\{r_1, r_2, \ldots, r_m\}$, and a hypothesis sentence H, the task maps them to ENTAIL (E), CONTRADICT (C) or NEUTRAL (N).

We can denote the mapping as follows.

$$f(\mathrm{T}, \mathrm{H}) \rightarrow \mathrm{y} \tag{7.1}$$

where, $\mathrm{y} \in \{\mathrm{E}, \mathrm{N}, \mathrm{C}\}$. For example, for the tabular premise in Table 7.1, the model should predict *E*, *C*, and *N* for the hypotheses *H1*, *H2*, and *H3*, respectively.

7.3.2 Trustworthy Tabular Inference

Trustworthy Tabular Inference is a table reasoning problem that seeks not just the NLI label, but also relevant evidence from the input table that supports the label prediction. We use T^R, a *subset* of T, to denote the relevant rows or evidence. Then, the task is defined as follows.

$$f(\mathrm{T}, \mathrm{H}) \rightarrow \{\mathrm{T}^R, \mathrm{y}\} \tag{7.2}$$

In our example table, this task will also indicate the evidence rows T^R of *Producer* and *Length* for hypothesis *H1*, *Recorded* for hypothesis *H2*, and *Released* and *Recorded* for hypothesis *H3*.

While the notion of evidence is well-defined for the ENTAIL and CONTRADICT labels, the NEUTRAL label requires explanation. To decide on the NEUTRAL label, one must first search for relevant rows (if any), i.e., identify evidence in the premise tables. In fact, this is a causally correct sequential approach. Indeed, INFOTABS has multiple neutral hypotheses that are partly entailed by the table; if any part of a hypothesis contradicts the table, then the inference label should be CONTRADICT. For example, in our example table, the premise table indicates that the album was recorded in 1978, emphasizing the importance of the *Recorded* row for the hypothesis H2. For NEUTRAL examples, we refer to any such pertinent rows as evidence.

7.3.3 Dataset Details

There are several datasets for tabular NLI: TabFact, INFOTABS, and the SemEval'21 Task 9 [226] and the FEVEROUS'21 shared task [5] datasets. We use the INFOTABS data in this work. It contains finer-grained annotation (e.g., TabFact lacks NEUTRAL hypotheses) and more complex reasoning than the others[3].

The dataset consists of 23,738 premise-hypothesis pairs collected via crowdsourcing

[3]As per [84], 33% of examples in INFOTABS involve multiple rows. The dataset covers all the reasoning types present in the Glue [270] and SuperGlue [269] benchmarks.

on Amazon MTurk. The tabular premises are based on 2,540 Wikipedia Infoboxes representing twelve diverse domains, and the hypotheses are short statements paired with NLI labels. All tables contain a *title* followed by two columns (cf. Table 7.1); the left columns are *keys* and the right ones are *values*).

In addition to the train and development sets, the data includes multiple test sets, some of which are adversarial: α_1 represents a standard test set that is both topically and lexically similar to the training data; α_2 hypotheses are designed to be lexically adversarial[4]; and α_3 tables are drawn from topics unavailable in the training set. The dev and test set, comprising of 7200 table-hypothesis pairs, were recently extended with crowdsourced evidence rows [83]. As one of our contributions, we describe the evidence rows annotation for the training set in the next Section 7.4.

7.4 Crowdsource Evidence Extraction

This section describes the process of using Amazon MTurk to annotate evidence rows for the 16,538 premise-hypothesis pairs that make the training set of INFOTABS. We followed the protocol of [83]: one table and three distinct hypotheses formed a HIT. For each of the hypotheses, five annotators would select the evidence rows. We divide the tasks equally into 110 batches, each batch having 51 HITs each having three examples. To reduce bias induced by a link between the NLI label and row selection, we do not reveal the labels to the annotators.

Since many hypothesis sentences (especially those with neutral labels) require out-of-table information for inference, we introduced the option to choose out-of-table (OOT) pseudo rows, which are highlighted only when the hypothesis requires information that is not common (i.e. common sense) and missing from the table. To reduce any possible bias due to unintended associations between the NLI label and the row selections (e.g., using OOT for neutral examples), we avoid showing inference labels to the annotators[5].

Human Annotation Quality Control: To assess an annotator, we compare their annotations with the majority consensus of other annotators' (four) annotations. We perform

[4]i.e. minimally perturbing hypothesis to flipped ENTAIL to CONTRADICT label and vice-versa.

[5]Because of the random sequence and unbalanced nature, each of the three hypothesis sentences can have any NLI label, i.e., in total $3^3 = 27$ possibilities.

this comparison at two levels: (a) **local-consensus-score** on the most recent batch, and (b) **cumulative-consensus-score** on all batches annotated thus far.We use these consensus scores to temporarily (local-consensus-score) or permanently (cumulative score) block the poor annotators from the task. We also review the annotations manually and provide feedback with more detailed instructions and personalized examples for annotators who were making mistakes due to ambiguity in the task. We give incentives to annotators who received high consensus scores. As in previous work, we removed certain annotators' annotations that have a poor consensus score (cumulative score) and published a second validation HIT to double-check each data point if necessary.

In total, we collected 81,282 annotations from 90 distinct annotators. Overall, twenty five annotators completed over 1000 tasks, corresponding to 87.75 % of the examples, indicating a tail distribution with the annotations. Overall, 16,248 training set table-hypothesis pairs were successfully labeled with the evidence rows[6]. On average, we obtain 89.49% F1-score with equal precision and recall for annotation agreement when compared with majority vote. Furthermore, 85% examples have an F1-score of >80 %, and 62% examples have an F1-score of >90 %. Around 60% examples have either perfect (100%) precision or recall, and 42% have both. Table 7.2 reports the Fleiss' Kappa score with annotation percentage. The average Kappa score is 0.79 with standard deviation of 0.23[7].

7.4.1 Choice of Semi-Structured Data

The rows of an Infobox table are semantically distinct, though all connected to the title entity. Each row can be considered a separate and uniquely distinct source of information about the title entity. Because of this property, the problem of evidence extraction is well-formed as relevant row selection. The same is not valid for unstructured text, whose units of information may be tokens, phrases, sentences or entire paragraphs, and is typically unavailable [79, 172, 221, 292].

---

[6]We exclude certain example pairings from our training sets since they could not achieve satisfactory agreement after adding more annotators or have label imbalance issues i.e. more the required number of neutrals.

[7]We also manually examined hypothesis phrases that signal relevant rows. See Appendix D.2 for details.

7.5 Trustworthy Tabular Inference

Trustworthy inference has an intrinsic sequential causal structure: extract evidence first, then predict the inference label using the extracted evidence data, knowledge/common sense, and perhaps formal reasoning [93, 192].[8] To operationalize this intuition, we chose a two-stage sequential approach which consists of an evidence extraction followed by the NLI classification, as shown in Figure 7.1.

Notation: The function f in Eq. 7.2 can be rewritten with functions g and h, $f(.) = g(.)$, $h \circ g(.)$, as

$$f(\mathrm{T}, \mathrm{H}) = \{g(\mathrm{T}, \mathrm{H}) \, , h\left(g(\mathrm{T}, \mathrm{H}), \mathrm{H}\right)\} \tag{7.3}$$

Here, g extracts the evidence rows T^R subset of T, and h uses the extracted evidence T^R and the hypothesis H to predict the inference label y, as

$$\begin{aligned} g(\mathrm{T}, \mathrm{H}) &\rightarrow \mathrm{T}^R \\ h(\mathrm{T}^R, \mathrm{H}) &\rightarrow \mathrm{y} \end{aligned} \tag{7.4}$$

To obtain f, we need to define the functions g and h, and a flexible representation of a semi-structured table T. To represent a table T, we use the **B**etter **P**aragraph **R**epresentation (BPR) heuristic of [182]. BPR uses hand-crafted rules based on the table category and entity type's of the row *values* (e.g., boolean and date) to convert each row to a sentence, consisting of table title, key and values. This representation outperforms the original "*para*" representation technique of [84].

We explore unsupervised (7.5.1) and supervised (7.5.2) methods for the evidence row extractor g.

7.5.1 Unsupervised Evidence Extraction

The unsupervised approaches extract Top-K rows are based on relevance scores, where K is a hyper-parameter. We use the cosine similarity between the row and the hypothesis sentence representations to score rows. We study three ways to define relevance described next.

[8] See more details discussion in 7.8.

7.5.1.1 Using static embeddings

Inspired by the **D**istracting **R**ow **R**emoval (**DRR**) heuristic of [182], we propose **DRR (Re-Rank + Top-S$_\tau$)**, which uses fastText [115, 168] based static embeddings to measure sentence similarity. We employ three modifications to improve DRR.

Re-rank (δ): We observed that the raw similarity scores (i.e., using only fastText) for some valid evidence rows could be low, despite exact word-level lexical matching with the row's *key* and *values*. We augmented the scores by δ for each exact match to incentivize precise matches.

Sparse extraction (S): For most instances, the number of relevant rows (K) is much lower than the total number of rows (m); most examples have only one or two relevant rows. We constrained the *sparsity* in the extraction by capping the value of K to S $\ll$ m.

Dynamic selection (τ): We use a threshold τ to select rows dynamically Top-K$_\tau$ based on the hypothesis, rather than always selecting fixed K r ows. We only select rows whose similarity (after Re-Ranking) to the hypothesis sentence representations is greater than a threshold τ. We adopt this strategy because (a) the number of rows in the premise table can vary across examples, and (b) different hypotheses may require a differing number of evidence rows.

7.5.1.2 Using word alignments

This approach consists of two parts (a) aligning rows and hypothesis words, and (b) then computing cosine similarity between the aligned words. Specifically, we use the SimAlign [106] method for word-level alignment. SimAlign uses static and contextualized embeddings without parallel training data to get word alignments. Among the approaches explored by SimAlign, we use the **Match (mwmf)** method, which uses **m**aximum-**w**eight **m**aximal matching in the bipartite weighted network formed by the word level similarity matrix. Our choice of this approach over the other greedy methods (Itermax and Argmax) is motivated by the fact that it finds the global optimum matching, while the other two do not. After alignment, we normalize the sum of cosine similarities of RoBERTa$_{\text{LARGE}}$ token embeddings[9] to derive the *relevance score*. Furthermore, because all rows use the same title,

[9]We use the average BPE token embeddings as the word embeddings.

we assign title matching terms zero weight. This chapter refers to this method as SimAlign (Match (mwmf)).

7.5.1.3 Using contextualised embeddings

The approach we saw in 7.5.1.2 defines row-hypothesis similarity using word alignments. As an alternative, we can directly compute similarities between the contextualised sentence embeddings of rows and the hypothesis. We explore two options here.

Sentence transformer: We use Sentence-BERT [23] and its variants [86, 256, 276], which use Siamese neural networks [34, 125]. We explore several pre-trained sentence transformers models[10] for sentence representation. These models differ in (a) the data used for pre-training, (b) the main model type and it size, and (c) the maximum sequence length.

SimCSE: SimCSE [71] uses a contrastive learning to train sentence embeddings in both unsupervised and supervised settings. The former is trained to take an input sentence and reconstruct it using standard dropout as noise. The latter uses example pairs from the MNLI dataset [280] with entailments serving as positive examples and contradiction serving as hard negatives for contrastive learning.

We give the row sentences directly to SimCSE to get their embeddings. To avoid misleading matches between the hypothesis tokens and those in the premise title, we swap the hypothesis title tokens with a single token title from another randomly selected table of the same category. We then use the cosine similarity between SimCSE sentence embeddings to compute the final relevance score. We again use the sparsity and dynamic selection as earlier. In the study, we refer to this method as SimCSE (Hypo-Title-Swap + Re-rank + Top-K^{τ}).

7.5.2 Supervised Evidence Extraction

The supervised evidence extraction procedure consists of three aspects: (a) Dataset construction, (b) Label balancing, and (c) Classifier training.

[10] https://www.sbert.net

7.5.2.1 Dataset construction

We use the annotated relevant row data (7.4) to construct a supervised extraction training dataset. Every row in the table, paired with the hypothesis, is associated with a binary label signifying whether the row is relevant or not. As before, we use the sentences from Better Paragraph Representation (BPR) [] to represent each row.

7.5.2.2 Label balancing

Our annotation, and the perturbation probing analysis of [83][11], show that the number of irrelevant rows can be much larger than the relevant ones for a table-hypothesis pair. Therefore, if we use all irrelevant rows from tables as negative examples, the resulting training set would be imbalanced, with about $6\times$ more irrelevant rows than relevant rows.

We investigate several label balancing strategies by sub-sampling irrelevant rows for training. We explore the following schemes: (a) taking all irrelevant rows from the table without sub-sampling (on average $6\times$ more irrelevant rows) referred to as **Without Sample($6\times$)**, (b) randomly sampling unrelated rowsin the same proportion as relevant rows, referred to as **Random Negative($1\times$)**, (c) using the unsupervised DRR (Re-Rank + Top-S_τ) method to pick the irrelevant rows that are most similar to the hypothesis, in equal proportion as the relevant rows, referred to as **Hard Negative($1\times$)**, and (d) same as (c), except picking three times as many irrelevant rows, referred to as **Hard Negative($3\times$)**[12].

7.5.2.3 Classifier training

We train a relevant-versus-irrelevant row classifier using $\text{RoBERTa}_{\text{LARGE}}$'s two sentence classifier. We use $\text{RoBERTa}_{\text{LARGE}}$ because of its superior performance over other models in preliminary experiments, and also the fact that it is also used for the NLI classifier.

7.5.3 Natural Language Inference

For the downstream NLI task, the function h is a two-sentence classifier whose inputs are T^R (the rows selected by g) and the hypothesis H. We use BPR to represent T^R as we did for the full table T. Since $|\text{T}^R| \ll |\text{T}|$, the extraction benefits larger tables (especially in

[11] Tabular probing using row deletion, row-value updation, row permutation, and row insertion.

[12] We explored other selection ratios too, take rows with rank till $5\times$, $2\times$, and $4\times$, but discovered that their performance is equivalent to (a), (b), and (c) respectively.

α_3 set) which exceed the model's token limit.

7.6 Experimental Evaluation

Our experiments assess the efficacy of evidence extraction (7.5) and its impact on the downstream NLI task by studying the following questions:

1. **RQ1:** What is the efficacy of unsupervised approaches for evidence extraction?(7.6.2)
2. **RQ2:** Is supervision beneficial? Is it helpful to use hard negatives from unsupervised approaches for supervised training? (7.6.2).
3. **RQ3:** Does evidence extraction enhance the downstream tabular inference task? (7.6.3)

7.6.1 Experimental Setup

First, we briefly summarize the models used in our experiments. We investigate both unsupervised (7.5.1) and supervised (7.5.2) evidence extraction methods. We use only the extracted evidence as the premise for the tabular inference task (7.5.3). We compare both tasks against human performance.

As baselines, we use the **W**ord **M**over **D**istance (WMD) of [84] and the original DRR [182] with Top-4 extracted evidence rows. For DRR (Re-Rank + Top-S^{τ}), which uses static embeddings, we set the sparsity parameter $S = 2$, and the dynamic row selection parameter $\tau = 1.0$. Our choice of S is based on the observation that in INFOTABS most (92%) instances have only one (54%) or two (38%) relevant rows. We set δ to 0.5 for all experiments.

For the Sentence Transformer, we used the *paraphrase-mpnet-base v2* model [23] which is a pre-trained with the *mpnet-base* architecture using several existing paraphrase datasets. This choice is based on performance on the development set.

Both the supervised and unsupervised SimCSE models use the same parameters as DRR (Re-Rank + Top-K_{τ}). We refer to the supervised and unsupervised variants as SimCSE-Supervised and SimCSE-Unsupervised respectively.

For the NLI task, we use the BPR representation over extracted evidence T^R with the $\text{RoBERTa}_{\text{LARGE}}$ two sentence classification model. We compare the following settings:(a) WMD Top-3 from [84], (b) No extraction i.e. using the full premise table with the "para" representation from [84], (c) DRR Top-4, (d) DRR (Re-Rank + Top-$2_{(\tau=1)}$) for training, development and test sets, (e) training a supervised classifier with a human oracle i.e.

annotated evidence extraction as discussed in 7.4, and using the best extraction model, i.e. supervised evidence extraction with Hard Negative (3×) for the test sets, and (f) the human oracle across the training, development, and test sets.

7.6.2 Results of Evidence Extraction

7.6.2.1 Unsupervised evidence extraction

For RQ1, Table 7.3 shows the performance of unsupervised methods. We see that the contextual embedding method, SimCSE-Supervised (Hypo-Title-Swap + Re-Rank + Top-$2_{(\tau=1)}$), performs the best. Among the static embedding cases, DRR (Re-Rank + Top-$2_{(\tau=1)}$) sees substantial performance improvement over the original DRR baseline. The alignment based approach using SimAlign underperforms, especially on the α_1 and α_2 test sets. However, its performance on the α_3 data, with out of domain and longer tables, is competitive to other methods.

Overall, the idea of using *Top-*S_τ, i.e., using the dynamic number of rows prediction and *Re-Rank* (exact-match based re-ranking) is beneficial. Previously used approaches such as DRR and WMD have low F1-score, because of poor precision. Using *Re-Rank* based on exact match improves the evidence extraction recall. Furthermore, introducing sparsity with *Top-*S_τ, i.e. considering only the Top-2 rows (S=2) and dynamic row selection ($\tau = 1$) substantially enhances evidence extraction precision. Furthermore, the zero weighting of title matches using the Hypo-Title-Swap heuristic, benefits contextualized embedding models such as SimCSE[13].

SimCSE-supervised (Hypo-Title-Swap + Re-Rank + Top-$2_{(\tau=1)}$) outperforms DRR (Re-Rank + Top-$2_{(\tau=1)}$) by 4.3% (α_1), 2.5% (α_2) and 5.4% (α_3) absolute score. Since the table domains and the NLI reasoning involved for α_1 and α_2 are similar, so is their evidence extraction performance. However, the performance of α_3, which contains out-of-domain and longer tables (an average of thirteen rows, versus nine rows in α_1 and α_2) is relatively worse. The unsupervised approaches are still 12.69% (α_1), 13.49% (α_2), and 19.81% (α_3) behind the human performance, highlighting the challenges of the task.

[13]For static embedding models, the effect of Hypo-Title-Swap was insignificant

7.6.2.2 Supervised evidence extraction

For RQ2, Table 7.4 shows the performance of the supervised relevant row extraction approaches that use binary classifiers trained with several sampling techniques for irrelevant rows. Overall, adding supervision is advantageous[14]. Furthermore, we observe that using the unsupervised DRR technique to extract challenging irrelevant rows, i.e., Hard Negative, is more effective than random sampling. Indeed, using random negative examples as the irrelevant rows performs the worst. Not sampling (6×) or using only one irrelevant row, namely Hard Negative (1×), also underperforms. We see that employing moderate sampling, i.e., Hard Negative (3×), performs best across all test sets.

The best supervised model with Hard Negative (3×) sampling improves evidence extraction performance by 8.7% (α_1), 10.8% (α_2), and 4.2% (α_3) absolute score over the best unsupervised model, namely SimCSE-Supervised (Hypo-Title-Swap + Re-Rank + Top-2$_{(\tau=1)}$).[15] The human oracle outperforms the best supervised model by 4.13% (α_1) and 2.65% (α_2) absolute scores—a smaller gap than the best unsupervised approach. We also observe that the supervision does not benefit the α_3 set much, where the performance gap to humans is still about 15.95% (only 3.80% improvement over unsupervised approach). We suspect this is because of the distributional changes in α_3 set noted earlier. This highlights directions for future improvement via domain adaptation.

7.6.3 Results of Natural Language Inference

For RQ3, we investigate how using only extracted evidence as a premise impacts the performance of the tabular NLI task. Table 7.5 shows the results. Compared to the baseline DRR, our unsupervised DRR (Re-Rank + Top-2$_{(\tau=1)}$) performs similarly for α_2, worse by 1.12% on α_1, and outperforms by 0.95% on α_3.

Using evidence extraction with the best supervised model, Hard Negative (3×), trained on human-extracted (Oracle) rows results in 2.68% (α_1), 3.93% (α_2), and 4.04% (α_3) improvements against DRR. Furthermore, using human extracted (Oracle) rows for both

---

[14] We investigate "How much supervision is adequate?" in 7.7.1.

[15] Although α_2 is adversarial owing to label flipping, rendering the NLI task more difficult, both α_1 and α_2 have instances with the same domain tables and hypotheses with similar reasoning types, making the relevant row extraction task equally challenging.

training and testing sets outperforms all models-based extraction methods. The human oracle based evidence extraction leads to largest performance improvements of 3.05% (α_1), 4.39% (α_2), and 6.67% (α_3) over DRR. Overall, these findings indicate that extracting evidence is beneficial for reasoning in tabular inference task.

Despite using human extracted (Oracle) rows for both training and testing, the NLI model still falls far behind human reasoning (Human NLI) [84]. This gap exists because, in addition to extracting evidence, the INFOTABS hypotheses require inference with the evidence involving common-sense and knowledge, which the NLI component does not adequately perform.

7.7 Evidence Extraction: Human versus Model

We perform an error analysis of how well our proposed supervised extraction model (Hard Negative(3x)) performs compared to the human annotators. The model makes two types of errors: a Type I error occurs when an evidence row is marked as irrelevant, whereas Type II error occurs when an irrelevant row is marked as evidence. A Type I error will reduce the model's precision for the extraction model, whereas a Type II error will decrease the model's recall. Type I errors are especially concerning for the downstream NLI task. Since mislabeled evidence rows will be absent from the extracted premise, necessary evidence will be omitted, leading to inaccurate entailment labels. On the other hand, with Type II errors, when an irrelevant row is labeled as evidence, the model has to deal with from extra noise in the premise. However, all the required evidence remains.

Table 7.6 shows a comparison of the supervised extraction (Hard Negative (3x)) approach with the ground truth human labels on all the three test sets for both error types. On the α_3 set, Type-I and Type-II errors are substantially higher than α_1 and α_2. This highlights the fact that on the α_3 set, the model disagrees with with humans the most. Furthermore, the ratio of Type-II over Type-I errors is much higher for α_3.

This indicates that the supervised extraction model marks many irrelevant rows as evidence (Type-II error) for α_3 set. The out-of-domain origin of α_3 tables, as well as their larger size, might be one explanation for this poor performance. Appendix §D.1 provides several examples of both types of errors.

7.7.1 Semi-Supervised Evidence Extraction

To investigate this, we use Hard Negative (3x) with RoBERTa$_{LARGE}$ model as our evidence extraction classifier, which is similar to the full supervision method. To simulate semi-supervision settings, we randomly sample 10%, 20%, 30%, 40%, and 50% example instances of the train set in an incremental fashion for model training, where we repeat the random samplings three times. Figures 7.2, 7.3, and 7.4 compare the average F1-score over three runs on the three test sets α_1, α_2 and α_3 respectively.

We discovered that adding *some* supervision had advantages over not having any supervision. However, we also find that 20% supervision is adequate for reasonably good evidence extraction with only $<$ 5% F1-score gap with full supervision. One key issue we observe is the lack of a visible trend due to significant variation produced by random data sub-sampling. It would be worthwhile to explore if this volatility could be reduced by strategic sampling using an unsupervised extraction model, an active learning framework, and strategic diversity maximizing sampling, which is left as future work.

7.8 Discussion

7.8.1 Why Sequential Prediction?

Our choice of the sequential paradigm is motivated by the observation that it enforces a causal structure. Of course, a joint or a multi-task model may make better predictions. However, these models ignore the causal relationship between evidence selection and label prediction [93, 192]. Ideally, each row is independent and, its relevance to the hypothesis can be determined on its own. In a joint or a multi-task model that exploits correlations across rows and the final label, *irrelevant rows* and the *NLI label*, can erroneously influence row selection.

7.8.2 Future Directions

Based on the observations and discussions, we identify the future directions as follows. (1) *Joint Causal Model.* To build a joint or a multi-task model that follows the causal reasoning structure, significant changes in model architecture are required. Such a model would first identify important rows and then use them for NLI predictions, but without risking spurious correlations. (2) *How much Supervision is Needed?* As evident from our

experiments, relevant row supervision improves the evidence extraction, especially on α_1 and α_2 sets compared to unsupervised extraction. But do we need full supervision for all examples? Is there any lower limit to supervision? We partially answered this question in the affirmative by training the evidence extraction model with limited supervision (semi-supervised setting) in Section 7.7.1, but a deeper analysis is needed to understand the limits. (3) *Improving Zero-shot Domain Performance.* As evident from §7.6.2, the evidence extraction performance of out-of-domain tables in α_3 needs further improvements, setting up a domain adaptation research question as future work. (4) Finally, inspired by [182], we may be able to add explicit knowledge to improve evidence extraction.

7.9 Conclusion and Future Work

In this chapter, we introduced the problem of *Trustworthy Tabular Inference*, where a reasoning model both extracts evidence from a table and predicts an inference label. We studied a two-stage approach, comprising an evidence extraction and an inference stage. We explored several unsupervised and supervised strategies for evidence extraction, several of which outperformed prior benchmarks. Finally, we showed that by using only extracted evidence as the premise, our approach outperforms previous baselines on the downstream tabular inference task.

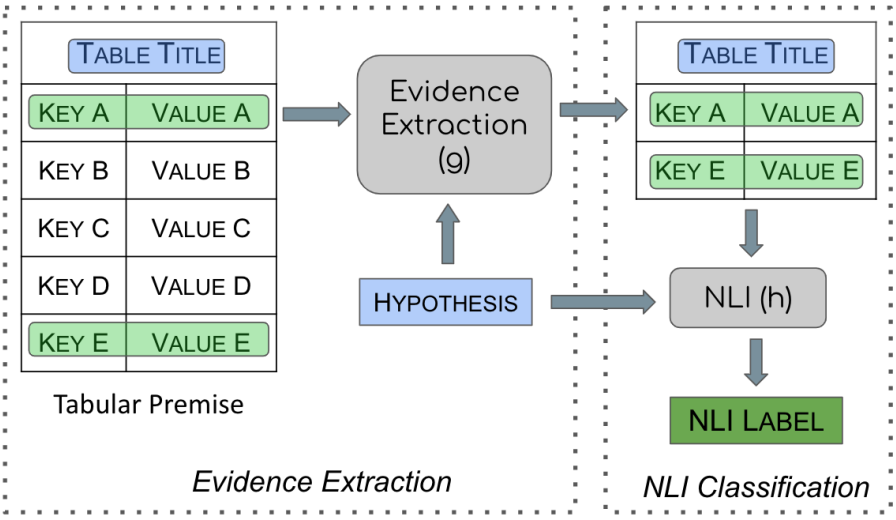


Figure 7.1: High level flowchart showing our approach for trustworthy tabular inference.

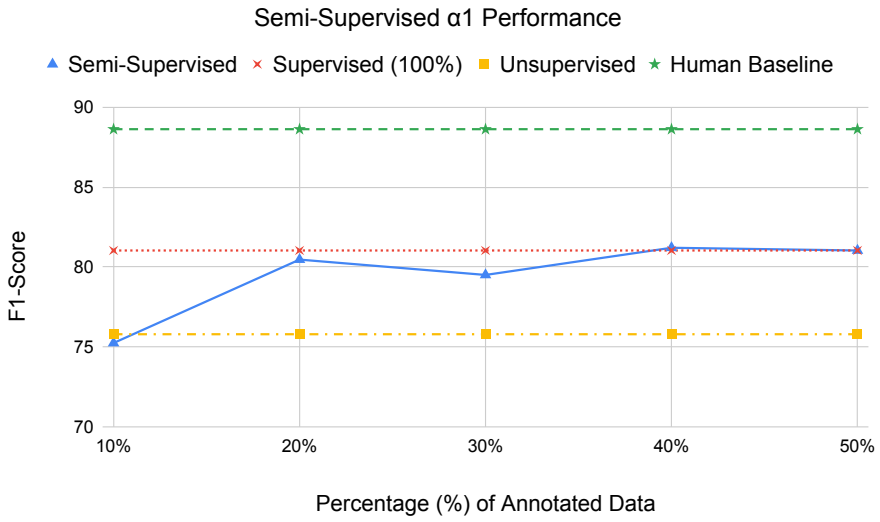


Figure 7.2: Extraction performance with limited supervision for α_1. All results are average of three random splits runs.

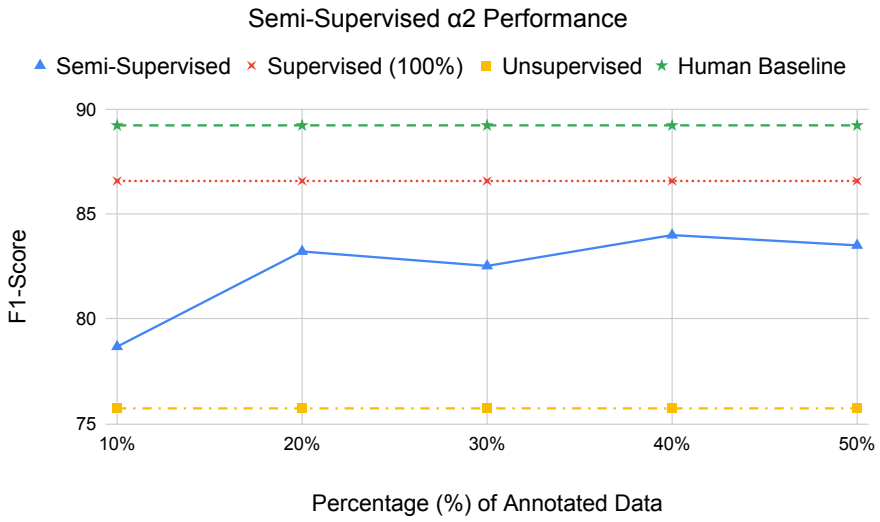


Figure 7.3: Extraction performance with limited supervision for α_2. All results are average of three random splits runs.

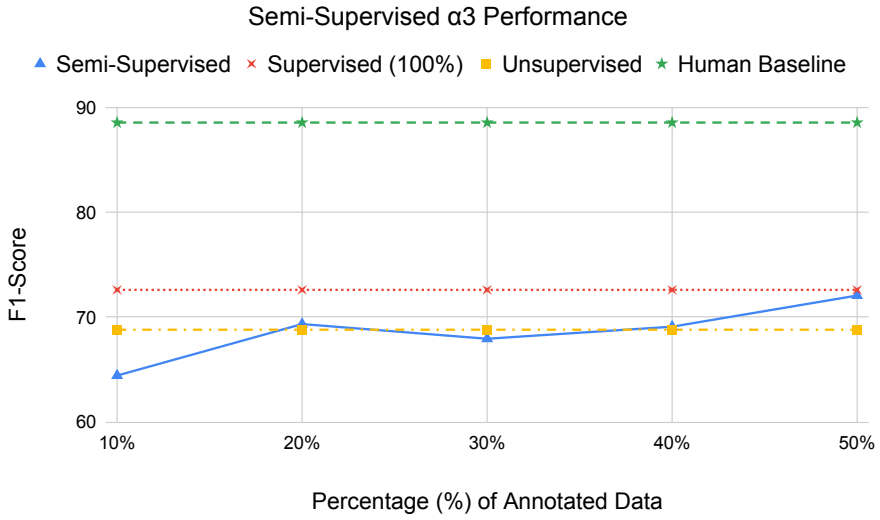


Figure 7.4: Extraction performance with limited supervision for α_3. All results are average of three random splits runs.

Table 7.1: A tabular premise example. The hypotheses H1 is entailed by it, H2 contradicts it, and H3, H4 are neutral i.e. neither entailed nor contradictory.

Breakfast in America	
Released	29 March 1979
Recorded	May–December 1978
Studio	The Village Recorder (Studio B) in Los Angeles
Genre	pop ; art rock ; soft rock
Length	46:06
Label	A&M
Producer	Peter Henderson, Supertramp

H1: Breakfast in America is a pop album with a length of 46 minutes.

H2: Breakfast in America was released at the end of 1979.

H3: Most of Breakfast in America was recorded in the last month of 1978.

H4: Breakfast in America has 6 tracks.

Table 7.2: Examples (%) for each Fleiss' Kappa score bucket.

Agreement	**Range**	**Percentage (%)**
Poor	< 0	0.27
Slight	0.01 – 0.20	1.61
Fair	0.21 – 0.40	5.69
Moderate	0.41 - 0.60	13.89
Substantial	0.61 - 0.80	22.92
Perfect	0.81 - 1.00	55.61
Overall	mean 0.79	s.t.d. 0.23

Table 7.3: F1-scores of the unsupervised evidence extraction methods.

Category	**Unsupervised Methods**	α_1	α_2	α_3
Baseline	WMD [84]	29.42	30.13	28.23
	DRR [182]	33.36	35.72	33.38
Static Embed.	DRR (Re-Rank + Top-2$_{(\tau=1)}$)	71.49	73.28	63.41
Alignment	SimAlign (Match (mwmf))	58.98	61.53	66.33
	Sentence-Transformer (paraphrase-mpnet-base-v2)	67.37	69.88	63.36
Contextualised	SimCSE-Unsupervised (Hypo-Title-Swap + Re-Rank + Top-2$_{(\tau=1)}$)	72.93	70.88	66.33
Embedding	SimCSE-Supervised (Hypo-Title-Swap + Re-Rank + Top-2$_{(\tau=1)}$)	**75.79**	**75.74**	**68.81**
Human	Oracle [83]	**88.62**	**89.23**	**88.56**

Table 7.4: F1-scores of supervised evidence extractors.

Sampling (Ratio)	α_1	α_2	α_3
Random Negative (1×)	69.42	71.94	54.12
Hard Negative (1×)	80.88	84.37	68.28
No Sampling (6×)	83.76	85.41	71.26
Hard Negative (3×)	**84.49**	**86.58**	**72.61**
Human Oracle	**88.62**	**89.23**	**88.56**

Table 7.5: Tabular NLI performance with the extracted relevant rows as the premise.

Category	**Evidence Extraction Train Set**	**Evidence Extraction Test Set**	α_1	α_2	α_3
Baseline	WMD [84]	WMD [84]	70.38	62.55	61.33
	No Extraction [84]	No Extraction [84]	74.88	65.55	64.94
	DRR [182]	DRR [182]	75.78	67.22	64.88
Unsupervised	DRR (Re-Rank + Top-2$_{(\tau=1)}$)	DRR (Re-Rank + Top-2$_{(\tau=1)}$)	74.66	67.38	65.83
Supervised	Oracle	Supervised (3× Hard Negative)	77.34	71.15	68.92
Human	Oracle	Oracle [83]	**78.83**	**71.61**	**71.55**
Human	Human NLI [84]	Human NLI [84]	**84.04**	**83.88**	**79.33**

Table 7.6: Type-I and Type-II error of best supervised evidence extraction model.

Test Set	**Type-I**	**Type-II**	**Ratio (II/I)**	**Total**
α_1	312	430	1.38	742
α_2	286	358	1.25	644
α_3	508	1053	2.07	1561

CHAPTER 8

TABULAR DATA AUGMENTATION

Adapted from D. Kumar, V. Gupta, S. Sharma, S. Zhang, *Realistic data augmentation framework for enhancing tabular reasoning*, in Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, Association for Computational Linguistics, pp. 4411–4429.

Human-generated datasets such as INFOTABS (Chapter 3) are limited in scale and thus insufficient for learning with large language models [51, 158]. Since curating these datasets requires expertise, huge annotation time, and expense, they cannot be scaled. Furthermore, in Chapter 3, we show that these datasets often suffer from annotation bias and spurious correlation problem [75, 87, 202].

In contrast, automatically generated data lacks diversity and have naive reasoning aspects. Recently, use of large language generation model [137, 208, 210] is also proposed for data generation [173, 190, 306]. Despite substantial improvement, these generation approaches still lack factuality, i.e., suffer hallucination, have poor facts coverage, and also suffer from token repetition (refer to Section 8.5 analysis). Recently, [25] shows that automatic tabular NLG frameworks cannot produce logical statements and provide only surface reasoning.

In this chapter, we address the above shortcomings, by utilising a semi-automatic framework that exploits the patterns in tabular structure for hypothesis generation. Specifically, this framework generates hypothesis templates transferable to similar tables since tables with similar categories, e.g., two athlete tables in Wikipedia, will share many common attributes. In Table 8.1 the premise table key attributes such as "Born", "Died", "Children" will soon be shared across other tables from the "Person" category. One can generate a template for tables in the Person category, such as <Person_Name> died before/after <Died:Year>. This template could be used to generate sentences as shown in Table 8.1

hypothesis H1 and $H1^C$. Furthermore, humans can utilize cell types (e.g., Date, Boolean) for generation templates. Recently, it has been shown that training on counterfactual data enhances model robustness [177, 212, 275]. Therefore, we also utilize the overlapping key pattern to create counterfactual tables. The complexity and diversity of the templates can be enforced via human annotators. Additionally, one can further enhance the diversity by automatic/manual paraphrasing [45] of the template or generated sentences.

To show the effectiveness of our proposed framework, we conduct a case study with INFOTABS dataset. INFOTABS is an entity-centric dataset for tabular inference, as shown in example Table 8.1. We extend the INFOTABS data (25K table-hypothesis pair) by creating AUTO-TNLI, which consists of 1,478,662 table-hypothesis pairs derived from 660 human written templates based on 134 unique table keys from 10,182 tables. For experiments, we utilize AUTO-TNLI in three ways (a.) as a standalone tabular inference dataset for benchmarking, (b.) as a potential augmentation dataset to enhance tabular reasoning on INFOTABS, i.e., the human-created data (c.) as evaluation set to assess model reasoning ability. We show that AUTO-TNLI is an effective data for benchmarking and data augmentation, especially in a limited supervision setting. Thus, this semi-automatic generation methodology has the potential to provide the best of both worlds (automatic and human generation). [1]

8.1 Contributions

We make the following contributions in this chapter:

- We propose a semi-automatic framework that exploits the patterns in tabular structure for hypothesis generation.
- We apply this framework to extend the INFOTABS [84] dataset and create a large-scale human-like synthetic data AUTO-TNLI that contains counterfactual entity-based tables.
- We conduct intensive experiments using AUTO-TNLI and demonstrate it helps benchmark and data augmentation, especially in a limited supervision setting.

This work was published at EMNLP 2022 Findings as [129]. We also constructed a framework for semi-automatically recasting existing tabular data to build tabular NLI

[1] The dataset and associated scripts, are available at https://autotnli.github.io.

instances from five database style tabular datasets that were initially intended for tasks like table2text creation, tabular Q/A, and semantic parsing. We demonstrate that recasted data could be used as evaluation benchmarks as well as augmentation data to enhance performance on tabular NLI tasks. Furthermore, we investigate the effectiveness of models trained on recasted data in the zero-shot scenario, and analyse trends in performance across different recasted datasets types. This work was also published at EMNLP 2022 Finding as [107].

8.2 Background

Synthetic creation of dataset has long been explored [119, 177, 225, 284]. For tabular NLI in particular, the datasets can be categorized into 1) Manually created datasets [84] with manually creates both hypothesis and premise, [26] manually creates the hypothesis while premise is automatically generated 2) Synthetically created semi-automatically generated datasets which completely automate data generation requires manual designing table-dependent context-free grammar (CFG) [58], or require logical forms to be annotated [25, 29, 177]. Several works such as [78, 87, 179, 187, 202, 266] have shown that models exploit spurious patterns in data. Similar to [84, 184, 298] authors investigate impacts of artifacts in dataset by creating adversarial test sets. However, semi-automatic systems requiring a CFG or logical forms contains reasoning which is often limited to certain types. Creating sentences that contain other reasonings (like lexical reasoning, knowledge, and common sense reasoning) is challenging using CFG and logical forms. Our work requires subject matter experts to create entity specific templates for each category which leads to creating sentences with multiple reasonings as well as complex reasonings.

8.3 Proposed Framework

Our framework includes four main components: (a.) Hypothesis Template Creation, (b.) Rational Counterfactual Table Creation, (c.) Paraphrasing of Premise Tables, and (d.) Automatic Table-Hypothesis Generation. Figure 8.1 shows the proposed framework of our approach.

8.3.1 Hypothesis Template Creation

For a particular category of tables (e.g., *movie*), the row attributes (i.e. keys) are mostly overlapping across all tables (e.g., *Length*, *Producer*, *Director*, and others). Therefore, this consistency across table benefits in writing table category specific **key-based rules** to create logical hypothesis sentences. We create such key-based rules for the following reasoning types: (a.) Temporal Reasoning, (b.) Numerical Reasoning, (c.) Spatial Reasoning, (d.) Common Sense Reasoning. Table 8.2 provides examples of logical rules used to create templates. We denote the category of a table as ***Category*** and the table row keys of as <*Key*>. In addition, each template is paraphrased to enhance lexical diversity.

Frequently, these key-based reasoning rules generalize effectively across several categories. For example, the temporal reasoning rule based on the date-time type could be minimally modified to work for < *Release Date*> f category ***Movies*** tables, as well as the <*Established Date*> of category ***University*** tables, in addition to the <*Born*> of category ***Person*** in Table 8.2. Additionally, reasoning rules can be expanded to incorporate multi-row entities from the same table's data, as illustrated in Table 8.2 for the numerical reasoning type. Other examples for the same are "The elevation range of <*City*> is <*HighestElevation*> − <*LowestElevation*>" for category ***City*** table, "<*SportName*> was held at <*location*> on <*date*>" for ***Sports*** category.

8.3.2 Rational Counterfactual Table Creation

We also construct counterfactual tables, as illustrated in Table 8.1, in which the values corresponding to the original table's keys are altered. This counterfactual table contains non-factual unreal information but is consistent, i.e., the table facts are not self contradictory. Language models trained on such counterfactual instances exhibit greater robustness [83, 177, 212, 275] and prevent the model from over-fitting its pre-learned knowledge. Benefiting model in grounding and examining the premise evidence as opposed to employing spurious correlation. To create counterfactual table, we modify an original table with k keys. For a given category, these k keys constitute a subset of the n possible unique keys ($n >= k$) for that category.

To construct a counterfactual table, we modify the original table in one or more of the following ways: (a.) keep the row as it without any change, (b.) adding new value to an

existing key, (c.) substituting the existing key-value with counter-factual data, (d.) deleting a particular key-value pair from the table, (e.) and add a missing new keys (i.e. a key from $(n - k)$), (f.) and adding a missing key row to the table. For creating counterfactual tables, for each row of existing, a subset of operation is selected at a random each with a pre-decided probability p (a hyper-parameter).

While creating these tables, we impose an essential key-specific constraints to ensure logical rational in the generated sentences. E.g. in the example Table 8.1, for the counterfactual table of *Janet Leigh (Counterfactual)*, the <*Born*> is kept similar to original of *Janet Leigh (Original)*, whereas <*Died*> has been substituted for another *Person* table, while ensuring the constraint BORN DATE < DEATH DATE i.e. Jan 13, 1994 (Died Date of Counterfactual Table) is after July 6, 1927 (Born Date of Counterfactual Table)). Without the following the constraint that BORN DATE < DEATH DATE, the table with become rationally incorrect or self contradictory.

8.3.3 Paraphrasing of Premise Tables

Lack of linguistic variety is a significant concern with grammar-based data generating methods. Therefore, we employ both automated and human paraphrase of premise tables to address diversity problem. For each key for of a given category, we create at least three to five simple paraphrased sentences of the key-specific template. E.g. for <*Alma Mater*> from ***Person***, possible paraphrases can be "<*PersonName*> earned his degree from <*AlmaMater*>", "<*PersonName*> is a graduate of <*AlmaMater*>", and "<*AlmaMater*> is a alma mater of <*PersonName*>". We observe that paraphrasing considerably increases the variability across instances.

8.3.4 Automatic Table-Hypothesis Generation

Once the templates are constructed as discussed in §8.3.1, they can be used to automatically fill in the blanks from the entries of the considered tables and create logically rational hypothesis sentences. To create contradictory sentences, we randomly select a value from a collection of key values shared by all tables to fill in the blanks. This replacement ensures that the key-specific constraints, such as the key-value type, are adhered to. Furthermore, we ensure that similar template with minimal token alteration is used to create entail contradict pair. This way of creating entail and contradiction statement pairs

with lexically overlapping tokens ensure that, future model trained on such data won't adhere spurious correlation from the tabular NLI data i.e. minimising the hypothesis bias problem [202]. For example, for movie "Ironman" movie with rows "Budget: $140 million" and "Box−office: $585.8 million", using the template <*Movie*> was a "hit if <*Box Office*> − <*Budget*> else flop" from example Table 8.2, one can generate hypothesis *entail*: "The movie Ironman was a hit" and *contradict*: "The movie Ironman was a flop".

8.4 The AUTO-TNLI Dataset

We apply our framework as described in §8.3 on an entity specific tabular inference dataset INFOTABS to construct AUTO-TNLI. INFOTABS [84] consists of pairs of NLI instances: a hypothesis statement grounded and inferred on premise table is extracted from Wikipedia Infobox table across multiple diverse categories. We construct the AUTO-TNLI dataset from a subset of the INFOTABS dataset (11 out of 13 total categories), which includes the original table plus five counterfactual tables corresponding to each original table, for a total of 10,182 tables. We retrieve 134 keys and 660 templates, which we utilize to generate 1,478,662 sentences. However, unlike INFOTABS, which contains 3 labels, ENTAIL, CONTRADICT and NEUTRAL, AUTO-TNLI contains only two labels ENTAIL and CONTRADICT.

As previously reported in the original INFOTABS paper (Chapter 3) by [84], annotators are biased towards specific keys over others. For example, for the category ***Company***, annotators would create more sentences for the key <*Founded by*> than for the key <*Website*>, resulting in an inherent hypothesis bias in the dataset. While creating the templates for AUTO-TNLI, we ensure that each key has a minimum of two hypotheses and a minimum of three (>3) premise paraphrases, which helps mitigate hypothesis bias. To address the inference class imbalance labeling issue, we construct approximately 1:1 ENTAIL to CONTRADICT the hypothesis.

We observe that most additional human labor required to build such sentences is spent on the set of key-specific rules and constraints that ensure the sentences are grammatically accurate. The counter-factual tabular data is logically consistent, i.e., not self-contradictory. Table 8.3 details the number of unique keys, the minimum/maximum/average number of keys, and the total number of sentences per table in AUTO-TNLI. As can be observed, the system generates a large amount of AUTO-TNLI data compared to limited INFOTABS while

using only a few human-constructed templates with key-specific rules and constraints.

We have chosen INFOTABS as it has three evaluation sets α_1, α_2, and α_3, in addition to the regular training and development sets. The α_1 set is lexically and topic-wise similar to the train set, and in α_2 the hypothesis is lexically adversarial to the train set. And in α_3 the tables are from topics not in the train set. Moreover, it has multiple reasoning types such as multi-row reasoning, entity type, negation, knowledge & common sense, etc. INFOTABS has all three labels ENTAIL, NEUTRAL, and CONTRADICT compared to just two labels in other datasets such as TABFACT.

8.4.1 Human Verification

To evaluate the quality and correctness of our data, we requested one of our human annotators (expert NLP Ph.D. Grad student) to assign a label to the generated hypothesis and select a score from 1 to 5 for the grammar and complexity of the sentences. The grammar score reflects how meaningful and lexically accurate the data is, and the complexity score indicates how difficult it is to label the hypothesis correctly. This was done for about 1300 premise-hypothesis pairs from AUTO-TNLI.

Analysis: As observed in Table 8.4, humans marked 99.5% of the data as correctly labeled and gave an average score of about 4.89 out of 5 for the grammatical accuracy of the sentences. The sentences in this data also received an average complexity score of 3.64 out of 5.

8.4.2 Reasoning for AUTO-TNLI

Our annotators classified all the distinct[2] templates from AUTO-TNLI into 14 reasoning types present in INFOTABS. Table 8.5 shows the individual reasoning type distribution across each category. The distribution statistics of reasoning types across each category is shown in Table 8.6. Table 8.7 shows that summary statistics across various reasoning types. Figure 8.2 gives distribution of extend of multiple reasoning in each individual examples.

Analysis: As we observe in Table 8.5 the cumulative frequency of reasoning types across each category is highest for Person followed by University and City and the average frequency of reasoning types across category is City followed by Person and Paint. In Table

[2]Templates for Provost and President are very similar so we don't consider them to be separate templates.

8.7 we see that the cumulative frequency of reasoning types across all categories is highest for simple lookup followed by lexical and numerical which have the same frequency.

8.5 Automatic Data Generation

Using GPT-J-6B, we generate 9–11 sentences per category. In total, we generated 110 sentences for 11 categories. We then classified each sentence into one of the following five classes: (a.) Correct - Both sentence and labels are correct. (b.) Factual error - Sentence is meaningful, but the label assigned to it is wrong. (c.) Overfit error - The same sentence as seen previously is generated without any lexical changes. (d.) Hallucination error - When knowledge from outside the tables provided is used to make a sentence. (e.) Repetition error - The same sentence is generated several times.

Analysis: As observed in Figure 8.3, out of all the 110 automatically generated hypothesis only 32.7% were *Correct* i.e. sentences were meaningful and the labels assigned to them are correct. Among the rest, about 52% had *Factual* errors in them and around 35% were *Hallucination* errors. This further demonstrates that a semi-automatic approach, such as ours, is preferable, as fully automated generating techniques are not reliable.

8.6 Experiments and Analysis

Overall, we address the following two research questions through our experiments:

RQ1: (a) Taking AUTO-TNLI as an evaluation set, how challenging is the TNLI task? (b) If fine-tuning on AUTO-TNLI beneficial?

RQ2: (a) Is it beneficial to use AUTO-TNLI as data augmentation for the TNLI task? (b) If so, will it also be useful in little supervision scenario?

8.6.1 Experiment Settings

We use RoBERTa$_{\text{BASE}}$ [158] (12-layer, 768-hidden, 12-heads, 125M parameters) and ALBERT-BASE [134] (12-layer, 768-hidden, 12-heads, 12M parameters) as our model for all of our experiments[3]. [182] shows data augmentation techniques that uses MNLI data for pre-training acts as implicit knowledge and enhances the model performance for INFOTABS.

[3]Due to the large scale of the AUTO-TNLI data, we favour BASE over LARGE models for conducting efficient experiments.

Therefore, we also explore implicit knowledge addition via data augmentation. In particular, we explored the following models: (a) RoBERTa$_{BASE}$ fine-tuned using the AUTO-TNLI dataset (b) RoBERTa$_{BASE}$, fine-tuned on the MNLI dataset and the AUTO-TNLI dataset (MNLI + AUTO-TNLI). Additionally, we also explore performance with RoBERTa$_{BASE}$ model fine-tuned sequential on all three MNLI, AUTO-TNLI and INFOTABS dataset. Due to limited space, we report all ALBERT [4] findings in Appendix E.

8.6.2 Using AUTO-TNLI as TNLI Dataset

In this section, we assess how challenging our AUTO-TNLI is compared to the INFOTABS datasets (i.e., RQ1).

8.6.2.1 Data splits

We first construct several train-dev-test splits of AUTO-TNLI such that: (a) splits have table from different domains (categories)[5] (b) splits have unique table row-keys, (c) premises in splits are lexically diverse. For the category-wise splits, we explore two ways (a) we divided categories randomly into train-dev-test. (b) we construct the splits after doing a cross-category performance analysis. In the cross-category analysis, we get all premise-hypothesis pairs generated from tables in one category (for example *person*) and train our model on this data. After this we test on premise-hypothesis pairs generated from all other categories (for example : *city*, *movie* etc.) one-by-one. We keep the difficult categories for the model to solve in the test set. This is accomplished by counting the number of times an category's accuracy falls below a specific threshold[6] and then selecting the entities with the highest frequency. We kept *book*, *paint*, *sports & events*, *food & drinks*, *album* in train-set, *person*, *movie*, *city* in dev-set and *organization*, *festival*, *university* in test-set.

For key-wise split, we explore two approaches (a) we divide the keys randomly into train-dev-test. (b) we decided splits based on the associated keys-values named entities type namely - *person*, *person type*, *skill*, *organization*, *quantity*, *date time*, *location*, *event*, *url*,

---

[4] Experiments on the development set showed that RoBERTa$_{BASE}$ outperforms other pre-trained language models. BERT$_{BASE}$ and ALBERT$_{BASE}$ reached an accuracy of 63% and 70.4% respectively

[5] by table domain/categories we refer to table entity types e.g. "Person", "Album", and others.

[6] We choose the threshold as 80%.

product after cross-entity performance analysis.. Similar to cross-category analysis above, here we get all premise-hypothesis pairs corresponding to keys in a single entity, for example let's say we choose the entity *person* and it includes the keys *written by*, *mayor*, *president* etc. then we get all premise-hypothesis pairs corresponding to these keys and train on them. After this we test on premise-hypothesis pairs corresponding to all other entities (for example : *persontype*, *skill*) one-by-one. We select the entities that are challenging for the model in the test set. This is accomplished by counting the number of times an entity's accuracy falls below a specific threshold[4] and then selecting the entities with the highest frequency. We kept the *url*, *event*, *person type*, *skill*, *product* in train-set, *quantity*, *other*, *person* in dev-set and *date time*, *organization*, *location* in test-set.

Finally, for the lexical diversity, we split via paraphrasing premise. Here too, we explore two different strategies (a) premises in train, dev, and test are not paraphrased, i.e., have similar templates. (b) premises in train, dev, and test are lexically paraphrased i.e. have distinct templates.

8.6.2.2 Using AUTO-TNLI only for evaluation (RQ1a)

We first explore how challenging is AUTO-TNLI is used as an evaluation benchmark dataset. To explore this, we compare the performance of pre-trained RoBERTa$_{\text{BASE}}$ model in four distinct settings, as follows (a.) without (w/o) fine-tuning, (b.) fine-tuned with INFOTABS, (c.) fine-tuned with MNLI, (d.) fine-tuned over both MNLI and INFOTABS in order and and evaluate it on AUTO-TNLI test-sets splits. For finetuning on MNLI and INFOTABS dataset, we only consider the ENTAIL and CONTRADICT while excluding the NEUTRAL label instances for training purposes.

Analysis: Table 8.8 shows a comparison of accuracy across all augmentation settings. The best is obtained when using both MNLI and INFOTABS for training. In the cases where we have used some fine-tuning with MNLI or INFOTABS we observed an average accuracy of 67.5%. Comparing this with zero-shot accuracy for INFOTABS where we observed accuracy of 58.9%, we can see that semi-automatically generated data is still challenging.

8.6.2.3 Using AUTO-TNLI for both training, and evaluation (RQ1b)

Next, we explore if providing supervision improves the performance on the AUTO-TNLI evaluation sets. To explore this, we compare pre-trained RoBERTa$_{\text{BASE}}$ model performance in two distinct settings, where we fine-tune on train set (a.) of AUTO-TNLI, (b.) of both MNLI and AUTO-TNLI in order and evaluate on AUTO-TNLI test-sets. Here too, we exclude the NEUTRAL label instances from MNLI.

Analysis: Table 8.8 shows a performance (accuracy) comparison across all augmentation settings. For all splits except paraphrasing, RoBERTa$_{\text{BASE}}$ achieves an average 80% accuracy. It shows that our semi-automated dataset AUTO-TNLI is as challenging as INFOTABS [84], which has an average accuracy of 70% across all splits and is manually human-generated and is one-tenth the size of AUTO-TNLI. Pre-finetuning with MNLI as augmented data (i.e., implicit knowledge) only improves the performance by 2%. Table 8.9 shows the category wise analysis of the results. Identical findings were also seen with ALBERT$_{\text{BASE}}$ model, c.f. Appendix E.

8.6.2.4 Cross-category analysis

We analyze how the semi-automatic data created performs across categories, i.e., training on one category and evaluating on the rest. This gave an idea of how training on data from one category generalizes over the rest. In Table 8.10, we have shown the accuracy when our model is trained on the categories written in rows and evaluated on the categories given in the columns.

Analysis: Here we observed that except some categories such as *Sports & Events*, *Album* and *City* the cross category accuracy is pretty high among the rest. *Album* seems to be quite a hard category with all categories giving a low cross-category accuracy when evaluated on it. *City* gave a challenging test set when trained on *Sport & Events*. *University* is the toughest test set for *Album*. When used as a test-set, *City* gave the least accuracy against *Sports & Events*, *Album* gives the least accuracy against *Paint*, *University* gave the least accuracy against *Sports & Events* and for the rest *Album* gave the least accuracy.

8.6.2.5 Cross-entity analysis

We analyze how the semi-automatic data created performs across entities, i.e., training on one entity and evaluating on the rest. This gave an idea of how training on data from one category generalizes over the rest. In Table 8.11, we have shown the accuracy when our model is trained on the entity written in rows and evaluated on the entities given in the columns.

Analysis: Here we observed that *Date & Time* is quite a tough test-set for most entities. *Quantity* is a tough test-set for *Skill* and *URL*. For *Skill* and *Person Type* are tough test-sets for *Location* and *Quantity* respectively. When used as a test-set, *URL* gave the lowest accuracy against *Person Type*, *Quantity* gave the lowest accuracy against *URL* and for the rest the *URL* gave the least accuracy.

8.6.3 Using AUTO-TNLI for Data Augmentation

We explore if AUTO-TNLI can be used as an augmentation dataset for INFOTABS (i.e. RQ2). Since INFOTABS include all three ENTAIL, NEUTRAL and CONTRADICT labels, where as AUTO-TNLI include only ENTAIL and CONTRADICT labels, we explore the inference task as a two-stage classification problem. In first stage, we train a RoBERTa$_{\text{BASE}}$ classification model to predicts whether a hypothesis is NEUTRAL versus NON-NEUTRAL (either ENTAIL or CONTRADICT). In second stage, we fine-tune a separate RoBERTa$_{\text{BASE}}$ model to further classify the NON-NEUTRAL prediction instances from stage one into ENTAIL or CONTRADICT label. Figure 8.4 illustrates the two-stage classification approach.

8.6.3.1 Comparison models

For first-stage we consider two training strategies: (a.) only train on INFOTABS, (b.) pre-finetune on both MNLI followed by training on INFOTABS. We consider multiple data augmentation technique for second stage training where we augment (a.) **Orig:** the AUTO-TNLI without counterfactual table instances, (b.) **Orig +Count:** AUTO-TNLI including counterfactual table instances[7], (c.) **MNLI +Orig:** both MNLI and AUTO-TNLI without counterfactual table instances, (d.) **MNLI +Orig +Count:** both MNLI and AUTO-TNLI including counterfactual table instances. Additionally, we compare all above methods

[7]We take five counterfactual tables for each original table.

with (e.) **No Aug** i.e. the approach where we do not augment any additional data.

8.6.3.2 Evaluation sets

We utilize the INFOTABS test sets, which include all three inference labels for evaluation. In addition to standard development and a test split (α_1), INFOTABS also has two adversarial test splits, namely α_2 and α_3. E.g. in the example Table 8.1 if hypothesis sentence *Janet Leigh was born* before *1940* is ENTAIL, then in α_2 after perturbation the instance became *Janet Leigh was born* after *1940* with label as CONTRADICT. The test set α_3 is a zero-shot evaluation set consisting of premise tables from different domains with minimal key overlaps with the training set premise tables. To better handle α_2 and α_3 test-sets, we include a counterfactual table and hypothesis in AUTO-TNLI.

8.6.3.3 Supervision scenarios

We analyse the effect of using AUTO-TNLI as augmentation data for INFOTABS in two setting (a) **Complete Supervision** where we use complete INFOTABS training set for final fine-tuning (b) **Limited Supervision** where we use limited INFOTABS supervision for second stages. We explore using 0% (i.e. no fine-tune), 5%, 15% and 25% of INFOTABS training set for final fine-tuning.

8.6.4 Complete INFOTABS Supervision (RQ2a)

Table 8.12 shows a comparison of accuracy across all augmentation settings. In the **first case**, when the first stage is only trained on INFOTABS, we observe an improvement of 1.6% and 1.2% percentage in α_1 and α_3 test-set through direct AUTO-TNLI data augmentation base pre-finetuning (Orig+Count) in comparison with no augmentation i.e. direct INFOTABS fine-tuning. We didn't see any substantial improvement in α_2 performance. Fine-tuning with MNLI followed by AUTO-TNLI (with counterfactual tables) further improve the performance by 0.6%, 2.0%, and 0.45% on α_1, α_2, and α_3 respectively.

For **second case**, when the first stage is trained on both MNLI, followed by INFOTABS, we observe an improvement of 1.60% and 0.67% percentage in α_1 and α_3 test-set through direct AUTO-TNLI data augmentation base pre-finetuning (Orig+Count) in comparison with no augmentation i.e. direct INFOTABS fine-tuning. Here too, we didn't see any substantial improvement in α_2 performance. Finetuning with MNLI followed by AUTO-TNLI

(with counterfactual tables) further improve the performance by 1.44%, 1.94%, and 0.83% on α_1, α_2 and α_3 respectively. Identical findings were also seen with ALBERT$_{\text{BASE}}$ model, c.f. Appendix E.

8.6.4.1 Ablation analysis - independent stage-1 and stage-2 performance

We also did an ablation study to access the performance of individual RoBERTa$_{\text{BASE}}$ models of both stages. Table 8.13 shows the performance for stage one classifier i.e. NEUTRAL versus NON-NEUTRAL. We observe that adding MNLI data for augmentation substantially improves the performance by 1.89%, 2.28%, and 2.05% for α_1, α_2, and α_3, respectively.

Table 8.14 shows the comparison between all settings of stage-2. In stage-2 adding counterfactual tables improve the performance by 2.75% and 1.42% in α_2 and α_3 respectively. We didn't see any substantial improvement in α_2 performance. If we pre-finetune further with MNLI along with AUTO-TNLI we further get an improvement of 5.42%, 3.33% and 2% in α_1, α_2, and α_3 respectively. Identical findings were also seen with ALBERT$_{\text{BASE}}$ model, c.f. Appendix E.

8.6.5 Consistency Analysis across Augmentation

We perform a consistency analysis on three setting, namely No Augmentation, Orig + Count and MNLI + Orig + Count to obtain a better estimate of where pre-training with AUTO-TNLI helps improve performance in INFOTABS. In Figures 8.5, 8.6, 8.7, and 8.8 we have shown the consistency graphs on the 3 settings.

Analysis: We observe in Figures 8.5, 8.6, 8.7, and 8.8 that the model is more prone to classifying CONTRADICT as ENTAIL than the other way around in α_1 set and there is a significant improvement after pretraining with AUTO-TNLI.For α_2 and α_3 sets we can see a considerable improvement in ENTAIL being classified as CONTRADICT from pretraining on AUTO-TNLI. Pretraining on AUTO-TNLI always results in improvements overall.

8.6.6 Performance across Different Reasoning Types

We take the 160 pairs from development and α_3 test sets each, from INFOTABS, that have been categorised into 14 reasoning types to assess the impact of pre-training on

various reasoning types, namely (a) numerical reasoning, (b) co-reference, (c) multi-row reasoning, (d) knowledge and common sense, (e) simple lookup, (f) negation, (g) lexical reasoning, (h) entity type, (i) named entities, (j) temporal reasoning, (k) subjective/out-of-table, (l) quantification, (m) syntactic alternations, and (n) ellipsis. The frequency of ENTAIL and CONTRADICT pairs being correctly classified is shown in Table 8.15 and Table 8.16 respectively.

Analysis: In Table 8.15 we observe that 9 out of 14 times in development and 12 out of 14 times in α_3-test sets MNLI + Orig + Count perform best. In Table 8.16 we observe that 10 out of 14 times in development set Orig + Count perform best.

8.7 Limited INFOTABS Supervision (RQ2b)

In this setting, we analyse the effect of limiting INFOTABS supervision for the second stage i.e. ENTAIL versus CONTRADICT. We explore using 0% (i.e. no fine-tune), 5%, 15% and 25% of INFOTABS training set for fine-tuning. Table 8.17 shows the performance for every augmentation settings. The table report average result over three random samples from AUTO-TNLI. We observe that augmenting with AUTO-TNLI improve performance for all percentages. Furthermore, the improvement is much more substantial for lower than higher percentages. Here too, the best performance are obtained via fine-tuning with MNLI followed by AUTO-TNLI for all percentages. In Table 8.18 we have also shown the accuracy for the first stage of the 2-stage classifier in the limited supervision setting with and w/o MNLI augmentation.

Effects of AUTO-TNLI Augmentation: Since AUTO-TNLI only contains ENTAIL and CONTRADICT labels, to check how pretraining with AUTO-TNLI affects the results in the limited supervision setting we had to use the 2-stage classifier where (a.) No Augmentation in first stage. (b.) Augmentation with MNLI in first stage. In Tables 8.19 and 8.20, we present the combined stage performance on limited supervision both w and w/o MNLI pre-training. The first stage classifier is again used to classify NEUTRAL vs. NON-NEUTRAL.

Analysis: As we can see in both Table 8.19 and Table 8.20 that the best is obtained by similar models in either case, with the only difference being that augmenting the first stage with MNLI helps improve the accuracy across all cases.

8.8 Discussion

8.8.1 Why Counterfactual Table Generation?

Tabular dataset is inherently semi-structured. Therefore, each category table has a specific set of keys. This enables us to create key-specific templates based on the entity-types of keys [182], which could be applied to millions of tables of a given category. Furthermore, as explained in §8.4, the templates also generalize across keys with similar value types across categories. All this is only possible due to the semi-structured nature of tabular data. Using counterfactual tables equips the model with more linguistically comparable. But oppositely labeled data to learn from, guaranteeing that the model can learn beyond the superficial textual artifacts and so becomes more resilient as shown by [119, 212]. As a result, when counterfactual data is included in the AUTO-TNLI, we observe performance improvement throughout all experimental settings. This learning is further verified by the findings for better gains in α_2, which comprises instances of linguistically comparable but oppositely labeled data instances.

8.8.2 Why Semi-Automatic Approach?

By examining the two diametrically opposed frameworks, namely a Human and an Automatic Annotation Framework, we may see many issues with both. Manually created data is prohibitively expensive and demands much human effort, limiting the ability to develop large-scale databases. Additionally, humans have a propensity to establish artificial patterns when manually creating a dataset, such as not giving all keys the same importance (explained in §8.4). While autonomous data generation is computationally efficient, it has many limitations. e.g., the inability to generate linguistically complex sentences and the difficulty of incorporating reasoning into the dataset. Because most deep learning models perform better with more data, producing large-scale datasets at a reasonable cost is critical while retaining data quality. With this in mind, we presented a "semi-automatic" architecture with the following benefits: (a.) It simplifies the creation of large-scale datasets. Using only 660 templates, we can generate 1,478,662 premise-hypothesis pairings from around 10,182 tables. (b.) The framework may be reused with additional tabular data as long as the structure is preserved. (c.) It enables the creation of linguistically and lexically diverse datasets. (d.) As shown in §8.4, hypothesis bias can be minimized by establishing

an adequate number of diverse templates for all keys of each category. (e.) The premises have been paraphrased in three ways to bring the required lexical diversity.

8.9 Conclusion

We introduced a semi-automatic framework for generating data from tabular data. Using a template-based approach, we generate AUTO-TNLI. We utilized AUTO-TNLI for both TNLI evaluation and data augmentation. Our experiments demonstrate the effectiveness of AUTO-TNLI and, by implication, our framework, especially for adversarial settings. For the future work, we aim to involve the creation of additional lexically varied and robust datasets and investigate whether the addition of neutrals could improve these datasets.

8.10 Limitations

This work has focused on entity tables, where the tabular structure and knowledge patterns are straightforward. Nevertheless, our templates technique does not generate maybe true/maybe false statements, i.e., neutral statements, as they need enhanced common sense (e.g., subjective usage) and unmentioned entity knowledge, i.e., information beyond the premise table. It is unknown how to generate good templates automatically, such as using neural generation methods rather than leveraging expert domain knowledge. Also, how these manually curated templates work when applied with more complicated tables like nested and hierarchical tables is under-explored. Theoretically, we can generate an infinite number of premise-hypothesis pairs, but the marginal utility might hurt the notion. Additionally, the zero-shot capabilities for out-of-domain tables are limited by the presumption that tables in similar categories resemble keys.

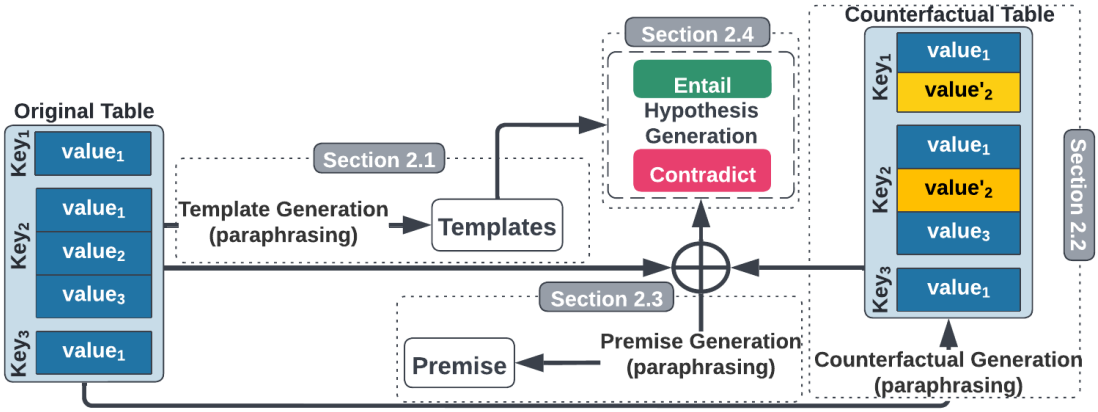


Figure 8.1: Our proposed framework. Yellow represents modified values in the counterfactual tables.

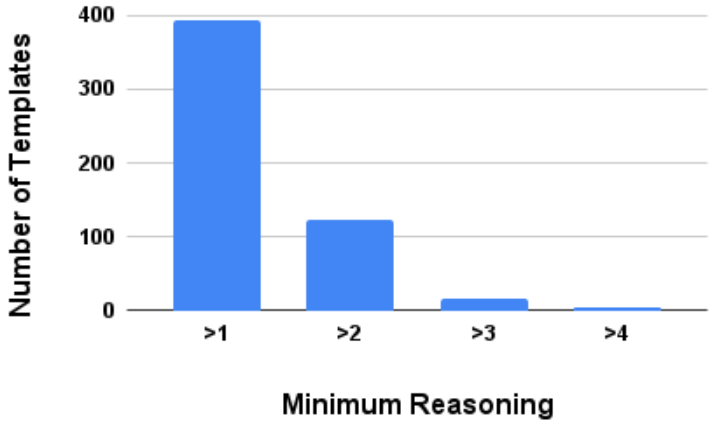


Figure 8.2: Cumulative frequency of templates across reasoning types in AUTO-TNLI.

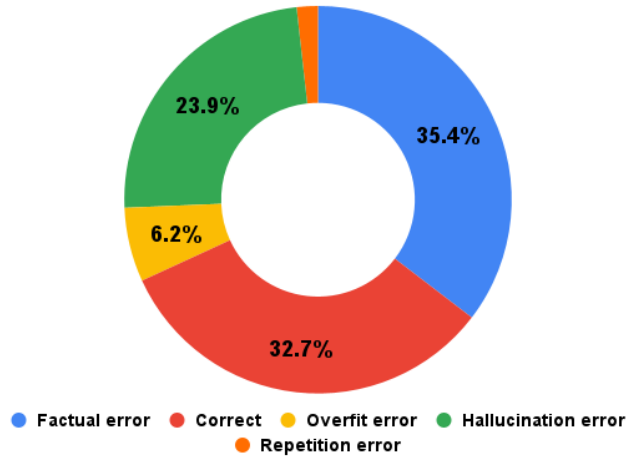


Figure 8.3: Percentage chart for automatic data generation correct and error labels.

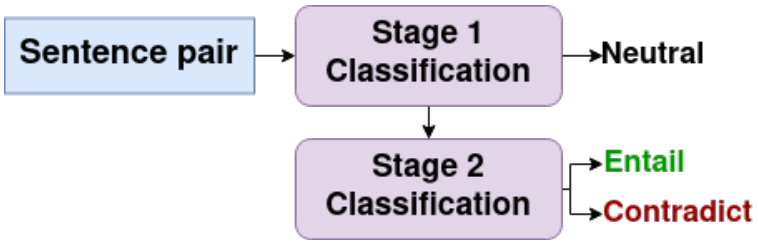


Figure 8.4: Two stage classification approach.

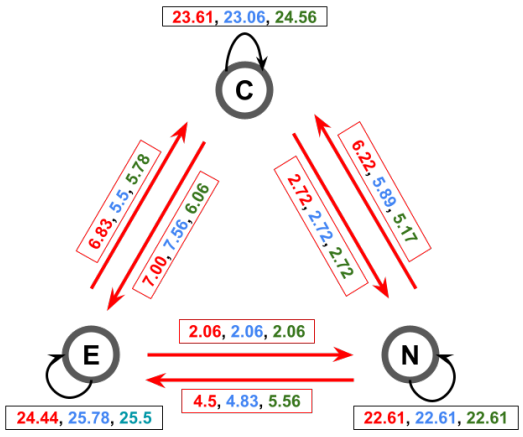


Figure 8.5: development consistency graphs. From top to bottom the values represent Red - No Augmentation, Blue - Orig+Counter, Green - MNLI+Orig+Counter.

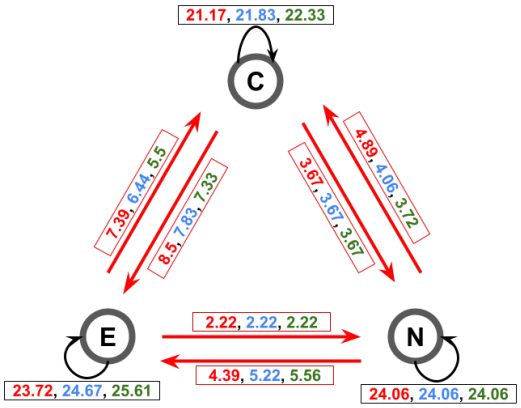


Figure 8.6: α_1 consistency graphs. Notation same as Figure 8.5.

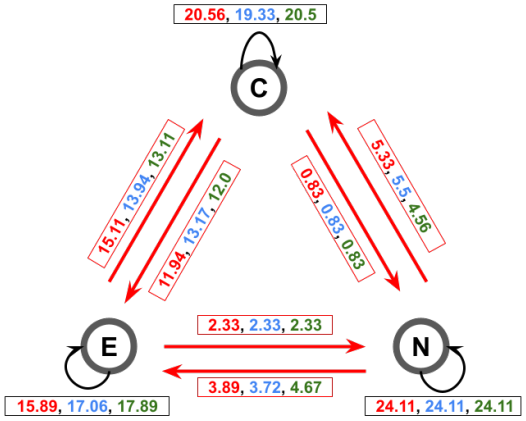


Figure 8.7: α_2 consistency graphs. Notation same as Figure 8.5.

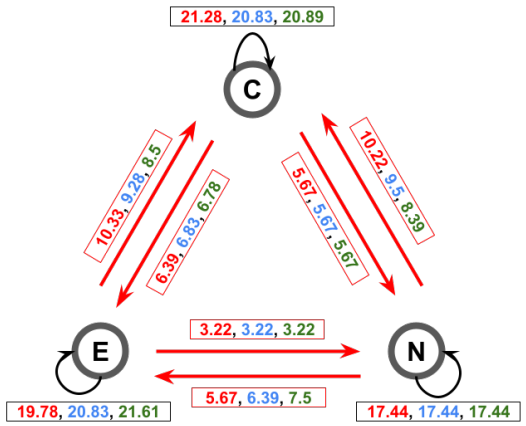


Figure 8.8: α_3 consistency graphs. Notations same as Figure 8.5.

Table 8.1: A example of an original and counterfactual table from the "Person" category. Here, we illustrate how multiple operations can be used to alter different keys. In addition, we have shown how the labels (E - Entail, C - Contradict) for a specific hypothesis can alter. In the "Janet Leigh" example table, the first column represents the keys (e.g. Born; Died) and the second column has the relevant values (e.g. July 6,1927; October 3, 2004 etc).

Janet Leigh (Original)			**Janet Leigh (Counter-Factual)**		
Born	July 6, 1927		**Born**	July 6, 1927	
Died	October 3, 2004		**Died**	January 13, 1994	
Children	Kelly Curtis; Jamie Lee Curtis		**Children**	Kelly Curtis	
Alma Mater	Stanford University		**Alma Mater**	University of California	
Occupation	None		**Occupation**	Scientist	
H1: Janet Leigh was born before 1940.		E	$H1^C$: Janet Leigh was born after 1915.		E
H2: The age of Janet Leigh is more than 70.		E	$H2^C$: The age of Janet Leigh is more than 70.		C
H3: Janet Leigh has 1 children		C	$H3^C$: Janet Leigh has more than 2 children.		C
H4: Janet Leigh graduated from Stanford		E	$H4^C$: Janet Leigh graduated from Stanford		C

Table 8.2: Rules and constraints are classified into specific areas of reasoning, as indicated in the table. A few examples of rules and constraints have been provided for each category. *<Died:Year>* indicates that the year value is extracted from *<Died>* , whereas *<Release1:Location>* indicates that the location is extracted from a single key-value pair in *<Release>*. KCS denote knowledge and common sense reasoning in this context.

Reasoning	**Category**	**Template-Rules**	**Table-Constraints**
Temporal	Person	*<Person>* was born in a leap year. *<Person>* died before/after *<Died:Year>*	Born Date $\leq$ Death Date
Numerical	Movie	*<Movie>* was a "hit if *<Box Office>* $-$ *<Budget>* else flop" *<Movie>* had a Box Office collection of *<BoxOffice>*	Budget ≥ 0
Spatial	Movie	*<Movie>* was released in *<Release1:Loc>*, *"X" months* before/after *<Release2:Location>*	Release1:Location $\neq$ Release2:Location
KCS	City	The governing of *<City>* is supervised by *<Mayor>* *<Mayor>* is an important local leader of *<City>*	Lowest Elevation $\leq$ Highest Elevation

Table 8.3: AUTO-TNLI statistics.

Statistic Metric	**Numbers**
Number of Unique Keys	134
Average number of keys per table	12.63
Average number of sentences per table	164.51

Table 8.4: Human verification statistics.

Statistic Metric	**Numbers**
Percentage of correct labels (%)	99.4
Average Grammar score (1-5)	4.89
Average Complexity score (1-5)	3.64

Table 8.5: Distribution of different reasoning types across all categories in AUTO-TNLI.

Statistics	City	Album	Person	Movie	Book	F&D	Org	Paint	Fest	S&E	Univ
numerical	19	7	28	24	16	8	19	3	14	9	5
co-reference	0	0	2	0	0	0	0	0	0	0	0
multi-row	4	0	15	5	0	7	6	7	1	1	6
KCS	21	2	45	9	3	0	24	5	0	5	27
temporal	1	6	31	5	1	0	1	4	3	6	2
syntactic-alt	23	0	6	10	2	8	6	2	14	4	28
simple-lookup	58	6	54	49	34	19	45	16	32	21	72
entity-type	0	0	54	0	0	0	1	4	0	3	1
ellipsis	0	0	20	0	0	0	1	0	0	1	0
subjective-oot	0	0	0	0	0	0	0	0	0	0	0
name-identity	0	0	6	0	0	0	3	0	0	0	0
lexical	25	0	7	20	19	3	30	2	13	6	27
quantification	13	5	19	11	11	3	8	1	10	11	3
negation	0	0	1	0	5	0	5	0	0	0	1

Table 8.6: Statistics of reasoning type distribution across the different categories in AUTO-TNLI.

Reasoning	City	Album	Person	Movie	Book	F&D	Org	Paint	Fest	S&E	Univ
No.	164	26	288	133	48	91	149	44	87	67	172
Avg.	2.52	1.37	2.25	1.77	1.78	1.86	1.99	2.2	1.74	1.68	2.12
Max	4	2	7	3	3	4	4	4	4	3	4

Table 8.7: Statistics of distribution of different reasoning types across all categories in AUTO-TNLI.

Reasoning	Average	Max	Min	Cumulative	Reasoning	Average	Max	Min	Cumulative
numerical	13.82	28	3	152	entity-type	5.73	54	0	63
co-reference	0.18	2	0	2	ellipsis	2	20	0	22
multi-row	4.73	15	0	52	subjective-oot	0	0	0	0
KCS	12.82	45	0	141	name-identity	0.82	6	0	9
temporal	5.45	31	0	60	lexical	13.82	30	0	152
syntactic-alt	9.36	28	0	103	quantification	8.64	19	1	95
simple-lookup	36.91	72	6	406	negation	1.091	5	0	12

Table 8.8: Accuracy with RoBERTa$_{BASE}$ model across several evaluation splits with / without fine-tuning on AUTO-TNLI. **bold** - represents max across rows i.e. best train/augmentation setting.

Training	Augmentation Strategy	Cat-Ran	Cross-Cat	Key	NoPara	Cross-Para	Entity
w/o AUTO-TNLI	w/o finetuning	50.00	49.64	50.17	49.77	49.75	49.78
	INFOTABS	66.17	63.86	**65.41**	65.15	65.12	63.66
	MNLI	67.15	64.95	64.79	65.33	65.33	62.2
	MNLI +INFOTABS	**69.28**	**65.9**	65.25	**66.41**	**66.39**	**65.02**
w AUTO-TNLI	Hypothesis-Only	53.74	55.1	58.82	66.47	66.86	56.36
	AUTO-TNLI	78.74	77.94	82.39	90.06	89.38	74.94
	MNLI +AUTO-TNLI	**83.82**	78.95	84.71	**91.17**	**90.57**	**77.66**
	MNLI +INFOTABS +AUTO-TNLI	83.62	**80.78**	**85.23**	90.98	90.03	77.19

Table 8.9: Category-wise results for AUTO-TNLI (**F&D**- Food & Drinks, **S&E** - Sports & Events)

Train-Data	City	Album	Person	Movie	Book	F&D	Org	Paint	Fest	S&E	Univ
Orig	78.32	67.81	92.45	97.12	96.31	92.27	92.44	98.93	87.44	82.53	85.59
Orig +Count	61.89	**68.26**	94.45	98.67	**98.72**	97.04	96.46	99.56	**93.73**	**95.68**	93.02
MNLI +Orig	**78.6**	68.12	92.89	97.74	97.21	93.19	93.06	99.36	88.12	84.18	87.03
++Count	62.32	68.01	**94.54**	**99.01**	98.46	**97.47**	**96.8**	**99.63**	93.66	95.08	**93.56**

Table 8.10: Cross-category analysis of our data. red - shows the least accuracy when trained on a category and evaluated on another. green - the least accuracy obtained when tested on a category and trained on the others. violet - intersection of the two cases above (**F&D**- Food & Drinks, **S&E** - Sports & Events)

Category	City	Album	Person	Movie	Book	F&D	Org	Paint	Fest	S&E	Univ
City	88.64	51.85	70.34	77.29	77	68.48	75.05	70.73	75.98	66.75	77.43
Album	52.92	79.35	65.2	60.28	57.38	65.75	59.16	53.48	58.8	55.75	52.9
Person	75.57	57.57	94.58	89.72	91.02	81.99	83.86	80.52	86.01	69.58	81.25
Movie	76.49	56.97	85.41	98.26	87.01	82.11	84.65	71.29	84.79	69.34	81.01
Book	54.03	53.37	76	77.69	97.84	78.68	76.81	73.51	64.94	71.62	53.76
F&D	61.79	56.72	80.67	83.24	87.55	95.82	80.46	76.49	74.61	68.71	58.03
Org	74.73	55.89	83.67	88.26	85.08	80.64	96.36	70.72	83.85	68.84	81.22
Paint	54.24	50.45	65.71	70.39	73.41	68.3	64.52	99	59.58	61.52	54.44
Fest	73.4	52.46	82.65	87.77	81.98	78.23	80.02	72.27	88.49	64.83	77.3
S&E	51.52	53.53	69.15	73.52	85.75	72.49	70.23	76.24	61.86	95.39	52.17
Univ	76.06	51.16	78.67	85.03	76.26	76.99	78.46	68.18	79.77	69.91	91.9

Table 8.11: Cross-entity analysis of our data. red - shows the least accuracy when trained on a entity and tested on another. green - the least accuracy obtained when tested on an entity and trained on the others. violet - intersection of the two cases above (**P&T**- Person Type, **D&T** - Date & Time)

Entity	Person	P&T	Skill	Org	Quantity	D&T	Location	Event	URL	Product	Other
Person	98.44	81.24	85.56	84.5	68.83	61.59	84.77	84.97	76.14	86.1	78.74
P&T	70.45	98.33	68.77	67.84	55.58	55.42	64.77	78.26	58.94	67.17	71.1
Skill	79.44	88.01	93.44	79.92	53.76	57.65	78.48	89.18	73.04	82.29	73.13
Org	92.36	87.33	86.58	95.62	63.56	58.03	87.19	87.12	84.09	86.9	81.29
Quantity	82.12	61.93	67.27	71.41	91.36	63.22	78.13	77	78.97	70.71	70.62
D&T	77.27	65.01	60.18	74.98	64.39	85.87	77.28	71.19	88.93	64.78	70.02
Location	88.32	76.32	86.3	83.18	68.89	62.31	94.43	81.57	83.69	79.98	75.75
Event	86.01	76.66	79.52	79.8	66.14	57.17	79.75	97.09	79.05	77.92	75.6
URL	61	56.27	58.42	60.88	51.61	55.02	62.68	60.56	95.25	56.07	55.09
Product	88.82	84.03	87.59	85.5	67.24	62.11	87.02	89.83	77.77	98.99	77.37
Other	83.39	84.98	80.82	78.24	62.44	58.29	76.97	86.74	69.98	82.78	93.88

Table 8.12: Accuracy of combine stage I i.e. NEUTRAL versus NON-NEUTRAL and stage II i.e. ENTAIL versus CONTRADICT classifiers (RoBERTa$_{BASE}$) across several data augmentation settings. Here, for stage one we also explore pre-fine tuning on MNLI data. **bold** - represents max across columns i.e. the best augmentation setting.

Stage 2: Entail versus Contradict					
Split	**No Augmentation**	**Orig**	**Orig+Count**	**MNLI+Orig**	**MNLI+Orig+Count**
Stage 1: INFOTABS					
dev	71.06	70.72	71.39	**72.28**	72.22
α_1	67.72	67.56	69.33	68.78	**69.89**
α_2	59.11	59.22	58.94	59.5	**61.28**
α_3	56.94	56.94	58.17	58.33	**58.61**
Stage 1: MNLI +INFOTABS					
dev	70.67	70.89	71.44	72.56	**72.67**
α_1	68.94	68.83	70.56	70.67	**72.00**
α_2	60.56	60.83	60.5	61.11	**62.50**
α_3	58.44	57.72	59.11	**60.06**	59.94

Table 8.13: Performance (accuracy) of stage one RoBERTa$_{BASE}$ (i.e. NEUTRAL versus NON-NEUTRAL) across several data augmentation settings. Here, No-Augmentation means INFOTABS, and MNLI means MNLI + INFOTABS. **bold** same as Table 8.12.

Test-split	**No Augmentation**	**MNLI**
dev	84.11	**84.50**
α_1	82.94	**84.83**
α_2	85.33	**87.61**
α_3	73.17	**75.22**

Table 8.14: Performance (accuracy) of stage two RoBERTa$_{BASE}$ (i.e. ENTAIL versus CONTRADICT) classifier across several data augmentation settings. **bold** same as Table 8.12.

Split	**No Augmentation**	**Orig**	**Orig+Count**	**MNLI+Orig**	**MNLI+Orig+Count**
dev	77.5	77.83	78.08	**80.75**	80.25
α_1	73.58	73.83	76.33	76.5	**79.00**
α_2	56.92	57.42	56.92	58.42	**60.25**
α_3	70.58	69.42	72	**73.08**	72.58

Table 8.15: Frequency of labels assigned as ENTAIL in each reasoning type across 3 settings and Gold labels for INFOTABS. **bold** - represents max across rows i.e. best train/augmentation setting.

	Human	No Aug	Orig +Count	MNLI +Orig +Count	Human	No Aug	Orig +Count	MNLI +Orig +Count
	Development set				α_3 set			
numerical	11	6	**8**	**8**	14	1	3	**5**
co-reference	8	**4**	**4**	3	5	2	2	**3**
multi-row	20	**13**	11	**13**	15	6	**8**	**8**
KCS	31	18	**21**	**21**	11	6	**9**	8
temporal	19	10	15	**16**	10	6	7	**8**
syntactic-alt	0	0	0	0	2	1	1	**2**
simple-lookup	3	**3**	**3**	**3**	8	**8**	7	**8**
entity-type	6	4	**5**	4	8	3	**6**	**6**
ellipsis	0	0	0	0	1	**0**	**0**	**0**
subjective-oot	6	3	**4**	**4**	2	**1**	**1**	**1**
name-id	2	**1**	**1**	**1**	1	**1**	**1**	**1**
lexical	5	**3**	**3**	**3**	3	2	**3**	**3**
quantification	4	1	**3**	**3**	2	**2**	**2**	**2**
negation	0	0	0	0	0	0	0	0

Table 8.16: Frequency of labels assigned as CONTRADICT in each reasoning type across 3 settings and Gold labels for INFOTABS. **bold** - represents max across rows i.e. best train/augmentation setting.

	Human	No Aug	Orig +Count	MNLI +Orig +Count	Human	No Aug	Orig +Count	MNLI +Orig +Count
	Development set				α_3 set			
numerical	7	**5**	**5**	**5**	14	**12**	10	7
co-reference	13	8	**10**	8	8	**6**	5	4
multi-row	17	**12**	**12**	**12**	12	**10**	8	8
KCS	24	15	**17**	16	17	**12**	**12**	**12**
temporal	25	15	**18**	15	16	**14**	11	12
syntactic-alt	0	0	0	0	0	0	0	0
simple-lookup	1	**0**	**0**	**0**	2	**2**	**2**	**2**
entity-type	6	3	**4**	**4**	9	**4**	3	1
ellipsis	0	0	0	0	0	0	0	0
subjective-oot	6	2	**3**	2	9	**5**	4	3
name-identity	1	**1**	**1**	**1**	0	0	0	0
lexical	4	**4**	3	**4**	8	**5**	4	3
quantification	6	3	**4**	**4**	4	**2**	1	**2**
negation	6	**6**	**6**	**6**	4	**3**	**3**	2

Table 8.17: Performance (accuracy) of RoBERTa$_{\text{BASE}}$ (i.e. ENTAIL versus CONTRADICT i.e. second stage) classifier with various data augmentation for limited supervision setting i.e. varying percentage of INFOTABS training data. The average standard deviation across 3 runs is 1.36 with range 0.5% to 3.5%. **bold** same as Table 8.12.

Tr(%)	No Augmentation	Orig	Orig+Count	MNLI+Orig	MNLI+Orig+Count
			Development set.		
0	50.25	59.58	52.58	**62.67**	60.75
5	65.31	69.92	69.86	70.81	**71.11**
10	67.53	72.08	69.83	**74.83**	73.42
15	69.47	71.69	73.61	**75.28**	74.42
20	71.28	73.61	72.47	**74.11**	**74.11**
25	70.21	72.88	74.54	**74.71**	74.63
			α_1 **set.**		
0	49.92	59.42	52.42	61.58	**62.33**
5	65.75	69.08	68.89	70.72	**70.92**
10	67.58	71.42	69	72.58	**74**
15	69.14	70.69	70.83	73.28	**74.25**
20	71.53	72.47	72.39	74.03	**74.61**
25	69.75	72.38	73.75	74.5	**75.13**
			α_2 **set.**		
0	50.17	59.00	59.75	61.17	**61.67**
5	43.81	54.92	53.53	56.25	**58.03**
10	47.92	54.08	54.5	**58.83**	56.75
15	47.31	54	53.03	56.89	**57.42**
20	49.17	54.03	54.44	**56.89**	55.75
25	49.79	56.33	55.25	**59**	58.42
			α_3 **set.**		
0	49.42	59.25	56.33	64.67	**63.92**
5	57.72	63.47	63.5	68.06	**68.14**
10	60.67	65.75	62.5	**71.58**	67.67
15	64.42	65.69	68.47	70.03	**71.11**
20	65.22	67.03	67.81	70.39	**71**
25	64.08	67.17	67.42	70.46	**70.92**

Table 8.18: First stage performance (accuracy) of RoBERTa$_{\text{BASE}}$ (i.e. NEUTRAL or NON-NEUTRAL) classifier with various data augmentation for limited supervision setting i.e. varying percentage of INFOTABS training data. The average standard deviation across 3 runs is 1.197 with range varying from 0% to 3.14%. **bold** same as Table 8.12.

Tr(%)	No Augmentation	MNLI	Tr(%)	No Augmentation	MNLI
	Development set			α_2 **set**	
0	**63.11**	59.28	0	64.61	**65.33**
5	75.75	**81.42**	5	78.17	**84.03**
10	76.86	**83.08**	10	79.86	**85.53**
15	78.92	**83.03**	15	81	**85.72**
20	78.83	**82.83**	20	81.58	**85.89**
25	78.92	**83.47**	25	81.81	**85.89**
	α_1 **set**			α_3 **set**	
0	**62.28**	58.5	0	**62.72**	56.06
5	76.94	**81.86**	5	69.42	**72.78**
10	77.11	**82.67**	10	69.17	**72.97**
15	79.22	**82.53**	15	70.22	**73.44**
20	78.53	**82.56**	20	67.86	**73.56**
25	78.92	**82.78**	25	68.81	**74.03**

Table 8.19: Both stage performance (accuracy) of RoBERTa$_{BASE}$ (i.e. ENTAIL, CONTRADICT or NEUTRAL) classifier with various data augmentation for limited supervision setting i.e. varying percentage of INFOTABS training data w/o MNLI pretraining for first stage. The average standard deviation across 3 runs is 0.98 with range varying from 0% to 4%. **bold** same as Table 8.12.

Tr(%)	No Augmentation	Orig	Orig+Count	MNLI+Orig	MNLI+Orig+Count
			Development set		
0	33.06	38.94	34.5	**39.56**	**39.56**
5	55.64	57.44	**59.11**	58.31	58.42
10	60	59.86	60.08	60.39	**61.36**
15	61.83	62.22	**63.44**	63.42	63.28
20	64.08	64.53	64.5	**65.31**	64.97
25	64.5	64.83	64.97	65.47	**66.11**
			α_1 **set**		
0	33	38.06	34.33	**40.28**	**40.28**
5	56.64	58.75	58.89	58.67	**59.03**
10	60.44	60.89	60.11	60.61	**61.31**
15	61.81	62.94	63.03	**64.33**	63.89
20	63.39	63.92	62.78	**64.11**	63.89
25	64.36	64.39	64.28	64.89	**65.69**
			α_2 **set**		
0	33.17	38.83	39.5	**40.5**	**40.5**
5	43.69	47.64	49.64	49.86	**51.44**
10	47.94	52.81	53.44	52.17	**54.83**
15	50.39	53.69	53.72	54.75	**54.94**
20	53.64	54.44	54.56	56.28	**56.72**
25	54.69	56	56.03	57.06	**57.94**
			α_3 **set**		
0	32.83	38.39	36.61	**41.61**	**41.61**
5	47.42	48.31	51.03	50.94	**52.31**
10	48	49.61	50.08	50.67	**52.86**
15	50.47	52.28	53.44	**53.86**	53.19
20	52.53	51.14	52.69	**54.17**	53.69
25	52.33	51.81	52.06	**54.25**	53.83

Table 8.20: Both stage performance (accuracy) of RoBERTa$_{BASE}$ (i.e. ENTAIL, CONTRADICT or NEUTRAL) classifier with various data augmentation for limited supervision setting i.e. varying percentage of INFOTABS training data with MNLI pretraining for first stage. The average standard deviation across 3 runs is 1.89 with range varying from 0% to 5.23%. **bold** same as Table 8.12.

Tr(%)	No Augmentation	Orig	Orig+Count	MNLI+Orig	MNLI+Orig+Count
Development set					
0	43.33	47	45.44	**47.94**	47.72
5	61.11	63.25	**64.81**	64.19	64.36
10	65.28	64.94	65.53	65.61	**66.78**
15	65.33	65.67	66.58	66.86	**66.89**
20	67.25	67.36	67.67	**68.78**	68.19
25	68.08	68.25	68.19	69.06	**69.56**
α_1 **set**					
0	42.06	47.94	45.78	**48.17**	48.11
5	61.83	64.06	64.08	64	**64.36**
10	65.64	**66.69**	64.92	65.86	66.44
15	64.39	65.72	65.58	**66.97**	66.61
20	66.44	66.97	65.69	**67.39**	66.78
25	67.69	68.03	67.61	68.17	**69.03**
α_2 **set**					
0	46.72	51.61	50.5	51.5	**52.06**
5	49.69	53.78	56.17	56.39	**57.67**
10	53	57.72	58.22	57.14	**60.22**
15	54.47	57.83	57.81	58.97	**59.14**
20	56.81	57.53	57.75	59.61	**60.11**
25	57.31	59.17	59.22	60.11	**61.08**
α_3 **set**					
0	39.72	43.33	42.33	**44.72**	44.17
5	50.64	51.67	54.14	54.56	**56.22**
10	52.08	53.39	54.11	55	**56.69**
15	53.72	55.58	56.67	**57.19**	56.75
20	55.5	54.61	55.94	**57.75**	57.69
25	56	55.39	55.89	**58.69**	57.92

CHAPTER 9

PATTERN EXPLOITED TRAINING

Adapted from A. Shankarampeta, V. Gupta, S. Zhang, *Enhancing tabular reasoning with pattern exploiting training*, in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Online, November 20-23, 2022, Association for Computational Linguistics, pp. 706–726.

Existing methods based on language models are ineffective for reasoning over semi-structured data, as discussed in Chapter 6. These models often ignore relevant rows and use spurious correlations in hypothesis or pre-training information for making inferences [83, 88, 104, 182, 202], as discussed in Chapter 7. Due to existing biases in human curated datasets [213, 310] with hypothesis having annotation artifacts [88], often models trained on such data lack generalizability and robustness [78], as discussed in Chapter 6. Furthermore, the absence of comprehensive test sets hinders robust model evaluation. Thus, evaluating models based only on accuracy does not reflect their reliability and robustness [175, 221].

In this Chapter, we investigate the current model's reasoning capability, particularly whether they can extract the right knowledge and correctly make rational inferences from that extracted knowledge. We focus on the task of tabular reasoning through table inference on INFOTABS [84] (Chapter 3). For instance, in Table 9.1, a model must filter out the relevant rows, i.e., extract knowledge, before applying the proper reasoning to categorize H1. Reasoning steps can be complex when involving numerical reasoning like count, sort, compare, arithmetic (H1: $46 < 50$), commonsense knowledge (H3: December occurs at the end of the year), and factual knowledge (H4: LA is short for Los Angeles).

It has been proven that LMs pre-trained without explicit supervision on a huge corpus of free web data implicitly incorporate several types of knowledge into their parameters [199]. For extracting this knowledge from language models (LM), various methods utilize

probing [95, 265], attention [105, 279], and prompting [200, 237] strategies. This internalized knowledge cannot be retrieved when fine-turning for a subsequent task. One explanation is that the objectives of pre-training and fine-tuning are vastly different. This variation in training objectives also diminishes the expected performance gains of the task, hence necessitating further pre-training on training data [58, 224, 284]. Therefore, reframing the subsequent task as a joint pre-training objective becomes essential. Hence, we reformulate the tabular NLI, i.e., our downstream task as a cloze-style problem, a.k.a, a mask language modeling (MLM) problem. For fine-tuning, we utilize the efficient Pattern-Exploiting Training (PET) technique [231, 232, 249]. PET entails establishing pairs of cloze question patterns and verbalizers that enable subsequent tasks to utilize the knowledge of the pre-trained language models. In addition, PET does not need model upgrades, such as adding more layers or parameters during pre-training.

Compared to direct fine-tuning-based techniques, i.e., training a classifier layer on top of LM, our method improved +8.1 and +25.8 on factual and relational knowledge evaluation tasks, respectively. On INFOTABS , a tabular inference dataset, our PET training approach outperforms +1.72 on α_1 (similar to dev), +2.11 on α_2 (adversarial set), and +2.55 on α_3 (zero-shot set), see Section 9.7) the existing baselines. This shows the effectiveness of our approach, especially on adversarial and out-of-domain challenging instances. Furthermore, we evaluate our improved model against instance perturbations to examine its robustness. These perturbations are generated by modifying existing INFOTABS instances, namely by changing names, numbers, places, phrases (paraphrasing), and characters (spelling errors). In addition, we also incorporated counterfactual instances (i.e., negation) to evaluate the model's robustness against pre-trained knowledge overfitting. The improvement in the counterfactual setting demonstrates that our approach benefits the model to ground better with premise table evidence. This work is published at AACL 2022 as [236].

9.1 Contributions

Our main contributions are the following [1]:

1. We propose a method for generating prompts for determining if current models can infer from knowledge.
2. We enhance the model's reasoning via prompt learning, i.e., PET, to extract knowledge from semi-structured tables.
3. Our experiments on INFOTABS show that our proposed approach preserves knowledge and improves performance on downstream NLI tasks. The results are robust when assessed on multiple curated adversarial test sets.

9.2 Background

9.2.1 Knowledge Incorporation and Evaluation

A line of works have been proposed to integrate knowledge into the LMs using pre-trained entity embeddings [199, 305], external memory [120, 162, 164], unstructured text [246, 284]. Several methods, including probing classifiers, have been proposed to extract and assess knowledge from LMs [95, 97, 265], attention visualization [105, 279], and prompting [111, 200, 237]. Many works have been published to study and create the prompts [156, 169, 206, 237].

9.2.2 Model Robustness

Many works proposed ways to evaluate robustness to noise, fairness, consistency, explanation, error analysis, and adversarial perturbations to test the model's robustness and reliability [78, 103, 155, 167, 179, 184, 218, 219, 220, 307]. [175] introduces a textual perturbation infrastructure that incorporates character- and word-level systematic perturbations to imitate real-world noise. [79] offered a toolbox to evaluate NLP systems on sub-populations, transformations, evaluation sets, and adversarial attacks.

9.3 Motivation

9.3.1 Case for Reasoning on Semi-Structured Data

Reasoning semi-structured data acquire skills such as arithmetic and commonsense, understanding the text types in the tabular cells, and aggregating information across nu-

[1] The dataset and associated scripts, are available at https://infoadapet.github.io/.

merous rows if necessary. For example, to judge the H1 in Table 9.1, the model needs to understand *"duration"* and *"length"* are the same in the context of the table, which is about a music album. Also, numerical reasoning is required to compare *"46:06" minutes"* is less than *"50 minutes"*. At the same time, the model should understand that the premise (table) is about a music album, so to classify the H1 model needs to understand the information present in 2 rows ({*"Genre"*, *"Length"*}) and perform numerical reasoning on top of that factual information.

9.3.2 Implicit Knowledge Is Required for Reasoning

For instance, for H3 in Table 9.1, the model needs to first extract the relevant row, i.e., *"Released"* row from the table, then compares the phrase *"end of 1979"* with the *"Released"* row value *"29 March 1979"* implicitly. The model needs to perform temporal reasoning to know that *"year 1979"* is correct. However, the month *"March"* is not the *"end of the year"*, but *"November"* or *"December"* is (implicit commonsense temporal knowledge). While previous works tried to incorporate knowledge via pre-training [58, 182]. In this work, we integrate knowledge and reasoning ability simultaneously using Pattern Exploiting Training [249]. This approach improves the existing knowledge and enhances reasoning compared to existing methods.

9.3.3 Robustness Is Critical for Model Evaluation

Tabular reasoning models typically fail on modest input modification, a.k.a. adversarial manipulation of inputs, highlighting the model's poor robustness and generalizability limit [83]. Thus, evaluating reasoning models on adversarial sets generated by minimal input perturbation becomes vital. As a result, we propose additional adversarial test sets, such as using character and word level perturbations to evaluate various aspects of model understanding and reasoning over tables. For example, if H1 (Table 9.1) is changed to *"Breakfast in Wales is a pop album with a duration of fewer than 50 minutes."* now the label of hypothesis H1 is changes from **entailment** to **neutral** since we do not know any information of *"Breakfast in Wales"* from Table 9.1. These minor input perturbations can alter the hypothesis' semantic interpretation. Idealistically, a robust model with superior reasoning ability should perform well on these input perturbed adversarial sets, as our technique also demonstrates.

9.4 Our Proposed Approach

In this section we describe our method to **(a)** evaluate pre-trained LM knowledge for tabular reasoning, **(b)** enhance model tabular reasoning capability using PET training, **(c)** and assess model robustness to input perturbations.

9.4.1 Evaluation of Pre-Training Knowledge

To examine how pre-training affects knowledge-based reasoning for tabular data, we focus on two types of knowledge (a.) factual knowledge (awareness of specific factual knowledge about entities), (b.) and relational knowledge (awareness of possible right relations between two distinct entities). For instance, in the sentence *"Breakfast in America was released on March 29, 1979"*, *"Breakfast in America"* and *"March 29, 1979"* are considered as factual knowledge, while their relationship term, i.e., *"released"* corresponds to relational knowledge.

We evaluate factual and relational knowledge in the language model before and after training for the downstream task like reasoning. In specific, we query the model using "fill-in-the-blank" cloze statements (a.k.a. prompts). As gauging knowledge using prompts is limited by how the prompts are constructed. We use part-of-speech tagging to detect nouns and verbs that are then used to mask names, numbers, and dates. These prompts are generated using hypotheses from the α_1, and dev sets as these sets have similar distribution as the training data [84]. We construct the prompts from both entailed and contradictory hypotheses. For prompts derived from entailed hypotheses, the model must predict the correct masked word, i.e., a term semantically equivalent to the word in the hypothesis. In contrast, for the prompts derived from contradicting hypotheses, the model should predict a semantically different term with the same entity type as the one mentioned in the hypothesis. To study the effect of the premise, we also query the model with the premise. To do this we modify the input as *premise* + *prompt*.

9.4.1.1 Prompts for factual knowledge evaluation

As most factual knowledge is contained in proper nouns and numbers, we randomly mask proper nouns or numbers in the hypothesis to generate a prompt and query the Language Model to fill the masked tokens. For example *"Duration of Breakfast in America*

is 46 minutes" (Table 9.1), *"Breakfast in America"*, *46* are the factual information present in the sentence and they are connected by *"duration"*. We randomly mask either *"Breakfast in America"* or *"46"* to generate prompt *"Duration of Breakfast in America is ¡mask¿ minutes"*. Occasionally, a masked term can be a number in numeric form (e.g., 2); however, the model predicted word form ("two"). We solved this issue by converting the predicted word into its numeric form or vice versa. E.g. *"Breakfast in America is produced by ¡mask¿ producers"*, where *¡mask¿ = two*.

9.4.1.2 Prompts for relational knowledge evaluation

Similar prompts are leveraged for relational knowledge. For example, to predict *¡mask¿ = released* for *"Breakfast in America was ¡mask¿ towards the end of 1979"*, the model needs to understand that *"Breakfast in America"* is a music album to predict *"released"* instead of *"eaten"* which is highly probable due the neighbor context term *"Breakfast"*. We also use WordNet [169] to discover synonyms for the masked term and see if the predicted word is among them.

9.4.2 Knowledge Incorporation for Reasoning

The issue of deducing inferences from tabular premises is similar to the typical NLI problem, except that the premises are tables rather than sentences. When evaluating the reasoning skills, we use a variety of representations of the tabular premise (see Section 9.5.3). We also study the effect of pretraining on an NLI task on INFOTABS.

9.4.2.1 Pattern-exploiting training

Using Pattern-Exploiting Training (PET) [231], NLU tasks are reformulated as cloze-style questions, and fine-tuning is performed using gradient-based methods. We use ADA PET (A Densely-supervised Approach to Pattern-Exploiting Training) [249], which increases supervision by separating the label token losses and applying a label-conditioned masked language modeling (MLM) to the entire input.

The input to the language model is converted into a cloze-style form with the pattern *<premise>* ? *<mask>*, *<hypothesis>*. The model is tasked to predict the masked word from the vocabulary. The model computes each token's probability as a softmax normalized overall tokens, allowing the logits of all vocabulary tokens to impact each likelihood,

similar to the regular MLM objective. While in PET, the masked word is forced to predict from the output space {*Yes, Maybe, No*} which are mapped to labels {*Entailment, Neutral, Contradiction*}. As a result, there will never be a gradient signal for non-label tokens. Inverting the query to the model to *"In light of the answer, what is the appropriate context?"* from *"What is the appropriate label based on the input?"* label conditioned mask language modeling is introduced by randomly masking out context tokens. If the label is "entail", during training, the model is obligated to predict the original token; however, if the label is "contradiction" or "neutral", the model is forced to ignore the original token.

9.4.2.2 Masked language modeling

ADAPET randomly masks tokens (RoBERTa style) from the context. Inspired by SpanBERT [114], ERNIE [246], we sample and mask the entire words based on pre-defined conditions. In Conditional Whole Word Masking (CWWM), we create a set of words S_w from a given sentence, and the POS of the words in that set must be from {"Adjective", "Adverb", "Noun, "Verb", "Proper Noun", "Adposition", "Numeral", "Coordinating Conjunction", "Subordinating Conjunction"}[2]. We sample words from the set S_w and mask all tokens matching the sampled word concurrently while maintaining the same overall masking rate. Figure 9.1 shows the training uses the two ADAPET components using MLM.

9.4.3 Robustness with Input Perturbations

We apply a range of character- and word-level perturbations to hypotheses to simulate circumstances where the input is slightly noisy or deviates from the training data distribution. We use TextAttack [176], NLP Checklist [221], and manual perturbations for generating the adversarial data. These adversarial sets will test the dependence of the model on word overlap, numerical comprehension, and hypothetical assertions. Refer to Tables 9.2 and 9.3 for examples.

Character-level perturbation employs perturbations such as introducing random characters, switching characters, removing a random character, and substituting a random character in the randomly selected word. This alteration does not impact the label of the hypothesis because it does not alter the sentence's meaning.

---

[2] https://universaldependencies.org/u/pos/

Location perturbation modifies the identified locations (countries, cities, and nationalities) in a sentence to another place specified in the location map. The NER model (TextAttack) identifies the location in a given sentence and replaces it with a sampled location from a dictionary. Here, cities are replaced with other cities and similar changes for countries. This perturbation transforms the entail clauses into contradictions but does not affect the original neutral and contradiction labels.

Name perturbation randomly replaces a person's name with the other one from a name list. This perturbation alters the label of every hypothesis into a neutral because the perturbed hypothesis and premise mention different persons.

Perturbing Numbers changes the entailed sentences into contradictions but does not affect the labels of neutral and contradictions. Contradictory statements remain contradictory because it is implausible that a randomly sampled number will be the actual number in the premise, making the hypothesis entailed.

Negation transforms entailment into a contradiction by negating the given sentence, keeping neutrals intact.

Paraphrasing paraphrases the given sentences without the loss of meaning using manual paraphrasing and Pegasus model[3]. Paraphrasing does not affect the inference label as it does not change the semantic meaning of the hypothesis.

Composition of Perturbations perturbs sentences by applying various distinct perturbations sequentially. E.g., in **num+para+name** we perturbed a sentence *"Supertramp, produced an album that was less than 60 minutes long"*, with premise Table 9.1 to *"Supertramp, produced an album that was less than 40 minutes long"* (number) then *"Supertramp released an album which lasted less than 40 minutes."* (paraphrase) then *"James released an album which lasted less than 40 minutes"* (name).

9.5 Experiments and Analysis

9.5.1 Dataset

Our experiments we use INFOTABS, a tabular inference dataset introduced by [84]. The dataset is diverse in terms of the tables domains, categories, and corresponding keys (entity types and forms) it contains, as illustrated in examples Table 9.1. In addition,

[3] https://biturl.top/MzQnMv

[84] reveals that inference on corresponding hypotheses requires extensive knowledge and commonsense reasoning ability. Given the premise table, hypothesis in the dataset is labeled as either an Entailment (E), Contradiction (C), or Neutral (N).

In addition to the conventional development set and test set (referred to as α_1), an adversarial test set (α_2) lexically equivalent to α_1 but with minor changes in the hypotheses to flip the entail-contradict label and a zero-shot cross-domain test set (α_3) containing large tables from other domains that are not in the training set are used for evaluation. For all of our experiments, we use the accuracy of classifying the labels as our primary metric for evaluation. The domain of tables in training sets and α_1,α_2 are similar. However, the training and fine-tuning tables are exclusive. Each of the test sets α_1, α_2, α_3 has 200 unique tables paired with 9 hypothesis sentences (3E, 3C, 3N), totalling 1800 table-hypothesis pairs. Table 9.4 depicts the statistics of perturbed sets from INFOTABS.

9.5.2 Models

We use the pre-trained RoBERTa-Large (RoBERTa$_L$) [158] language model from HuggingFace [281] for all of our investigations. We employ various configurations of language models to assess knowledge in two different cases. These configurations include RoBERTa$_L$, RoBERTa$_L$ finetuned on INFOTABS (RoBERTa$_L$+CLS), RoBERTa$_L$ trained for tabular inference using PET (ADAPET), and finetuning INFOTABS on ADAPET+CLS. Here we define finetuning as training a classifier head (CLS). We also investigate the effect of NLI pre-training using RoBERTa$_L$ pretrained on MNLI [280], and mixed dataset (mixNLI) containing ANLI+MNLI+SNLI+FeverNLI [4] [17, 183, 185]. All models are trained on 16538 table-hypothesis pairs (1740 tables) for 10 epochs with a 1e-5 learning rate.

9.5.3 Table Representation

We explored two ways to represent table (a.) *Table as paragraph* uses Better Paragraph Representation for table representation, (b.) and *Distracting Row Removal* prunes tables based on the similarity between hypothesis and tables rows. We investigated the pruning of top 4 (DRR@4) and top 8 (DRR@4) rows for our experiments. Both representation methods are adapted from [182].

---

[4] https://biturl.top/e6Vney

9.5.4 Results and Analysis

Our experiments answer the following questions:

1. **RQ1:** Can the large language model use pre-trained knowledge for reasoning? Does our adaptive training method enhance model reasoning?
2. **RQ2:** Does fine-tuning downstream tasks benefit model reasoning? Can our adaptive training benefit model via enhancing its reasoning knowledge?
3. **RQ3:** Is our adaptive method-based model robust to input perturbations? Can our method enhance model's semantic-syntactic comprehension?

9.5.4.1 Models knowledge evaluation

To answer RQ1, we evaluate the knowledge in the presence and absence of the premise using the Entail and Contradictory hypotheses, which are taken from the evidence in the premise tables. We do not use Neural statements as they may contain subjective and out-of-table information.

In all the settings (Tables 9.5 and 9.6) with and without premise, our model outperformed RoBERTa$_L$+CLS. The addition of the premise enhances model performance further. This can be ascribed to additional knowledge in the premise that our PET-trained model can leverage efficiently for reasoning. From Table 9.5, we observe that for all settings, our approach gave $\tilde{1}00\%$ improvement in relational knowledge evaluation compared to RoBERTa$_L$+CLS. Even training a classifier on top of ADAPET outperforms RoBERTa$_L$+CLS. We also evaluated on contradiction hypothesis to assess if the model can rightly identify false claims despite having correct entity types.

There is a significant difference between the Top 1 accuracy of premise+E and premise+C for factual knowledge evaluation as the model should not predict the masked token in the prompt from a contradiction statement, especially in factual prompts. And for relational knowledge, irrespective of the label of the hypothesis, the model should predict the masked token correctly if the model rightly understands the entity types of words in the sentence. In almost all the settings, our approach performs almost comparable to RoBERTa$_L$, and it even outperforms RoBERTa$_L$ in only Entail, and Premise+ Entail settings. Training a classifier on top of RoBERTa$_L$ decreases the performance knowledge evaluation but training a classifier head on top of ADAPET still tops RoBERTa$_L$+CLS, thus

demonstrating the benefits of our approach. A similar observation was reported with Top 5 accuracy (Table 9.6).

9.5.4.2 Knowledge incorporation for reasoning

To answer RQ2, we experiment with various premise representations of tables as paragraphs (BPR, DRR@4, DRR@8) (see Table 9.7). We observe that Roberta-Large with ADAPET improves performance in all premise representations except for α_3 with BPR compared to RoBERTa$_L$+CLS due to an increased number of keys in the tables (13.1 per table in α_3 when compared to 8.8 per table in α_1 and α_2). Results in Table 9.7 are the average accuracy of the models tested on multiple seeds.

We experiment with premise as a linearized table and compared our results with [84], see Table 9.8. Our proposed approach was able to outperform the baselines in [84] by a significant margin.

With ADAPET, we also improve performance using linearized table compared to [84] (+1.04 in α_1, +0.58 in α_2, +0.69 in α_3). ADAPET (token masking, no pre-training) tops RoBERTa$_L$+CLS in every premise representation and test split. +1.72 in α_1, +2.11 in α_2, +2.55 in α_3 with DRR@4. CWWM with ADAPET also outperformed RoBERTa$_L$+CLS. However, the performance of the two masking procedures is comparable for all test sets, even with the classifier setting.

We notice that the DRR@8 representation outperforms the best, especially in α_3 due to removing the irrelevant rows (+4.34 over BPR, +0.64 over DRR@4). The zero-shot test set α_3 which has a significant proportion of unseen keys (different domain tables) when compared to other test sets (number of unique keys intersection with train is 312, 273, 94 for α_1, α_2 and α_3 respectively) has seen a substantial improvement with the use of NLI pre-trained model. When compared to ADAPET (token masking, no pretraining), there has been an improvement of +2.13 units (no CLS) and +2.54 units (with CLS) with DRR@8 over no pre-training. We also observed that pre-training in more diverse data helps improve performance [6, 205]. Models which are pre-trained on mixNLI[4] outperformed MNLI pre-trained in almost every setting (+0.8 in α_1, +1.9 in α_2, +2.2 in α_3 with no CLS, DRR@8).

9.5.4.3 Robustness to input perturbation

To answer RQ3, we evaluate our model on several challenging input perturbations. The perturb test sets are generated using various character-level, and word-level perturbations are also tested with BPR, DRR@4, and DRR@8 table representations (see Table 9.9). To generate these sets, we applied perturbations on *dev*, and α_1 sets as the distribution of these sets are similar to the training set.

Except for the perturbations involving names, our method ADAPET (no pre-training) outperforms RoBERTa$_L$+CLS. We see the max improvement of ADAPET in the Negation (+4.4); this implies our model can handle counterfactual statements well. We observed that training a classifier head on top of ADAPET performed better with the adversarial sets involving multiple perturbations. In the challenge set with *number+paraphrase* all the ADAPET-based models outperformed RoBERTa$_L$+CLS by 2x times. We observed that using NLI pre-training also helps substantially improve the robustness. With the use of mixNLI and MNLI pre-trained weights, the performance of ADAPET-based models improved substantially compared to those without pre-training, even outperforming RoBERTa$_L$+CLS. From Table 9.9, it is clear that with hypotheses involving multiple perturbations, RoBERTa$_L$+CLS tends to perform more poorly compared to the ADAPET-based model. The performance on all perturb sets is much worse than that of the corresponding model on dev, α_1 sets. Improving the performance of these sets is crucial.

Qualitative Analysis of Perturbation Sets: On a randomly sampled subset containing 100 examples from each of the perturbation sets, we task a human evaluator to label them and give a score (out of 5) to the grammar of the hypotheses. For most cases, i.e., 11 out of 14, we observe a correct of $> 80\%$ indicating the correction of our adversarial tests. Furthermore, in half of the cases (7/14), the correctness score was above 95%. Grammar analysis shows that most sentences are highly grammatical, with an average score of 4.5/5.0. In the perturbation *"number+paraphrase"* we only observed 77% of label correctness. This could be due to changing numbers, followed by paraphrasing, which changed some contradiction hypotheses to neutral ones. A similar observation is also observed in *"number+char"* where numbers are modified in character perturbation. We also compare the models' performance on these sampled perturbed sets after human corrections in labels and grammar. We observed that the performance on these corrected sets is similar to the

generated perturbed sets, as in Table 9.10.

We also evaluate robustness with premise representation. In Tables 9.10 and 9.11 we show the performance of the model on the adversarial tests which are trained and tested with DRR@4, BPR representations of premise. We found the results are similar to the results in Table 9.9. Human correctness evaluation of the perturbation set is hown in Table 9.12. We also study the classification of Entailed and Contradictory hypotheses when the model is trained and tested on the data without any Neutral hypotheses, see Table 9.13. We found that DRR@4, DRR@8 representations of premise performs better that BPR because of the less distracting premise.

9.5.5 Error Analysis

When compared to Figure 9.2, in Figure 9.3, there is a substantial improvement in identifying NEUTRAL and CONTRADICTION, but there is also a confusion in identifying ENTAILMENT. Using the NLI-pre-trained model improves the detection of ENTAILMENT. A similar observation is also observed with using classifying layer (+CLS) (see Figures 9.3 and 9.4).

In Figure 9.5, we see the greatest inconsistency is with NEUTRAL being misidentified as ENTAILMENT across all models, and this is not that significant with using the classifying layer (+CLS) (see Figures 9.6, 9.7, and 9.8). Although with the classifying layer, there is increased confusion about CONTRADICTION being predicted as ENTAILMENT. Figure 9.9 represents the confusion Matrix between gold labels versus predictions of ADAPET(CWWM) and ADAPET(CWWM)+CLS model.

Table 9.14 shows a subset of the validation set labeled based on the different ways the model must think to put the hypothesis in the correct category. On average, all the ADAPET-based models perform similarly, but the human scores are better than the model we utilize. We observe that for certain reasoning types, such as Negation and Simple Look-up, neither humans nor the model arrives at the correct hypothesis, demonstrating the task's difficulty. For Numerical, Lexical, and Entity type reasoning, our model comes very close to human scores.

In Table 9.15, we observed that the City category on proposed models performs worse probably as a result of the engagement of more numeric and specific hypotheses com-

pared to the other categories, as well as longer average table size. Our models perform extremely well in identifying ENTAILMENT in Food & Drinks category because of their smaller table size on average and hypothesis requiring no external knowledge to reason as compared to CONTRADICTION. Our models also struggle in detecting NEUTRAL and CONTRADICTION in Organization category.

9.5.5.1 What did we learn?

Reformulating the NLI task as an MLM problem enabled the inclusion of premise table knowledge into Language Models (LM) for efficient reasoning. Using ADAPET, we have shown that knowledge can be retained and assimilated into reasoning tasks more effectively. ADAPET training also improves the model's ability to reason on downstream tasks. Similar observation is also observed in prior works [246, 284] where MLM is utilized to incorporate external knowledge, although the later require additional table based pre-training. Moreover, [83, 139] have shown that the LM utilizes spurious patterns to accomplish reasoning tasks. Our perturb sets study informed us that our ADAPET-based method is more robust than direct classification to semantic-syntactic alternations (see Section 9.6 for further discussions).

9.6 Further Discussion

Why table as a paragraph? A massive data corpus is used to pre-train the large language models. In contrast to semi-structured data, the bulk of pre-training data is unstructured. These models should, of course, perform better on unstructured data and struggle with semi-structured data. Tables in INFOTABS [84] are semi-structured in nature. These tables do not explicitly state the relationship between the keys and values; they can also have variable schemas. The album's overall duration is 46:06 minutes, according to the row with key Length and value 46:06. It is difficult to comprehend implicitly that "Length" refers to time length in minutes. Because of the absence of implicit information, a simple table linearization will not be sufficient. [84, 182] experimented with various forms of table representations. They found that representing tables as paragraphs gave better results and can leverage the advantage of pre-trained models datasets like MNLI for even better performance.

Why NLI task as cloze-style questions? While [88] showed MLM pre-training with unlabeled target data could further improve the performance on downstream tasks. [33] also showed that using MLM pre-training makes models robust to lexicon-level spurious features. [277] presented a methodology for analysis that connects the pre-training and downstream tasks to an underlying latent variable generative text model. They observed that prompt tuning achieves downstream assurances with less stringent non-degeneracy constraints than head tuning. By reformulating the NLI task as cloze style questions, we can use label conditioned MLM with prompt tuning, which resulted in a better performance on tabular reasoning on INFOTABS.

9.7 Conclusion

In this work, we have validated the effects of factual and relational knowledge in the language model via handcrafted prompts for tabular reasoning. Through prompt learning, i.e., Pattern-Exploiting Training, we extracted knowledge from semi-structured tables and further improved the model's reasoning capabilities. Our intensive experiments on the INFOTABS demonstrate that our approach can conserve knowledge and enhance tabular NLI performance. The conclusions hold up well when tested against carefully crafted adversarial test sets based on character and word-level perturbations.

9.7.1 Method Limitations

Entity tables are the focus of our solution. Its scalability in constructing prompts and other tables with different structures is limited by the idea that manually identified pattern from the specific dataset and template-based prompts. In addition, as not different from other NLP tasks, automatically detecting knowledge patterns and bridging patterns to prompts, especially for semi-structured tables, is under-explored. Furthermore, investigating prompting for sophisticated structured tables such as nested structures (e.g., lists inside tables), hierarchical tables (e.g., table inside a table), and multi-modal tables (pictures within table) will necessitate substantial effort.

9.7.2 Future Directions

We have identified the following future directions: (a.) *Designing better prompts for knowledge evaluation*: Our current prompts treat entail and contradictory statements as the

same while evaluating knowledge. In the presence of the premise, masking *Breakfast in America* in H3 (Table 9.1) and using that as an input model will predict Breakfast in America even though the hypothesis is a contradiction. We want to work on developing prompts label conditioned evaluation based on existing work on prompt engineering. [156]. (b.) *Improving Robustness:* While our models' performance on the challenging adversarial test sets is lower than benchmarks on INFOTABS , we do not know its reason. The created test sets may be challenging because they focus on phenomena that existing models cannot capture or exploit blind spots in a model's training set. Following the ideas of Inoculation by Fine-Tuning [155], we want to improve and assess the reasons behind the results in Table 9.9.

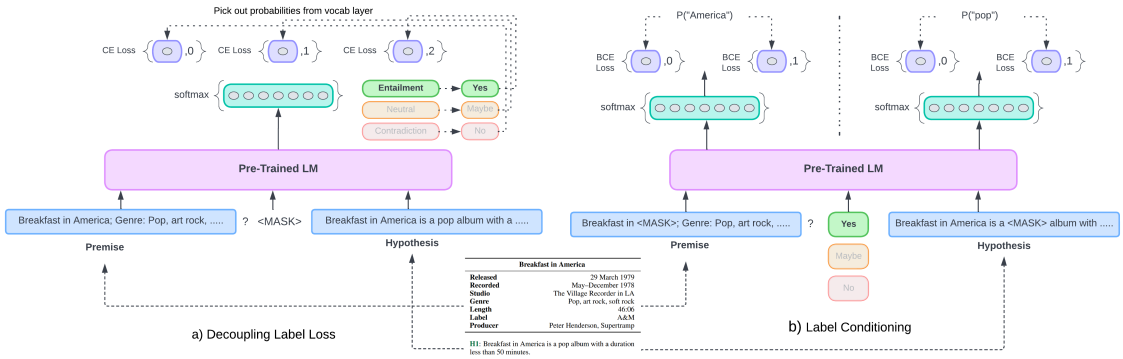


Figure 9.1: The training uses the two ADAPET components. Here, the blue boxes represent the task inputs (entailed, in this case) a) Decoupling Label Loss: Using the cross entropy loss across all labels, the model must predict the right and wrong labels at the masked-out position. b) Label Conditioning: The model should predict the original token at a randomly masked-out position if the input text has the entail label. Otherwise, not if the label is contradiction or neutral.

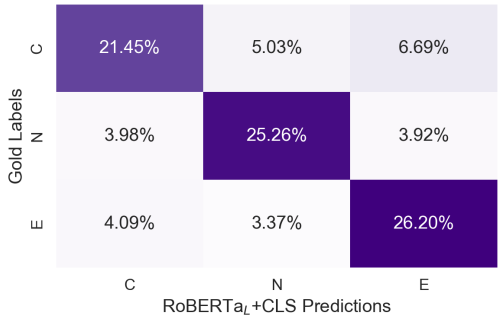


Figure 9.2: Confusion Matrix: Gold Labels vs predictions of RoBERTa$_L$+CLS.

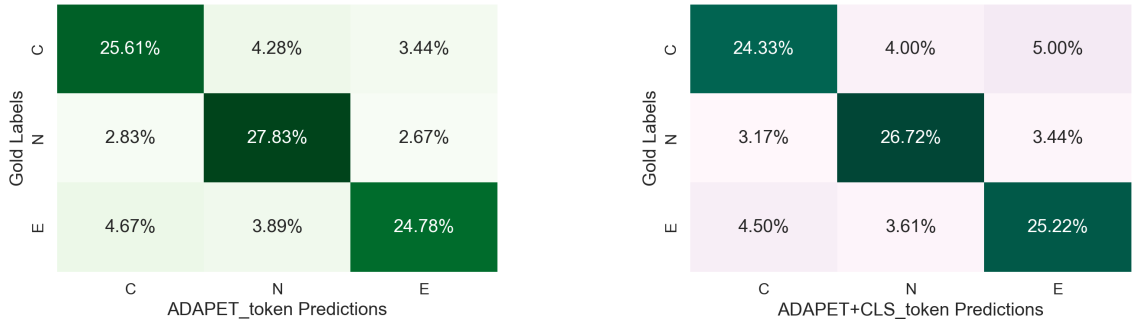


Figure 9.3: Confusion Matrix: Gold Labels versus predictions of ADAPET(token), ADAPET(token)+CLS.

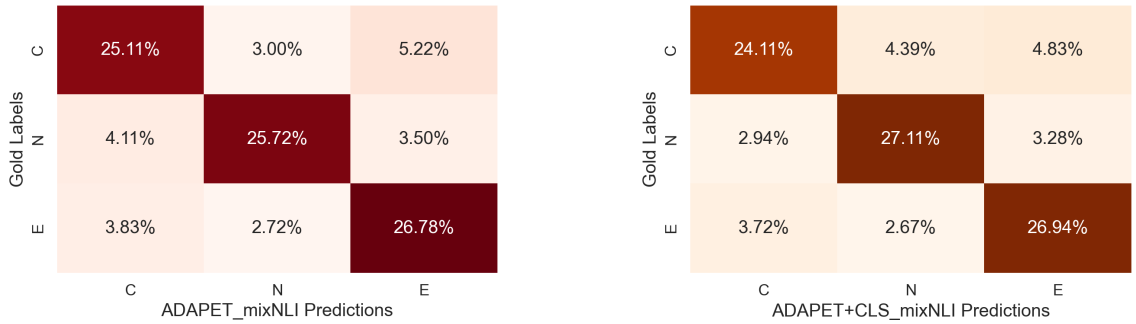


Figure 9.4: Confusion Matrix: Gold Labels versus predictions of ADAPET (pretrained mixNLI), ADAPET (pretrained mixNLI)+CLS.

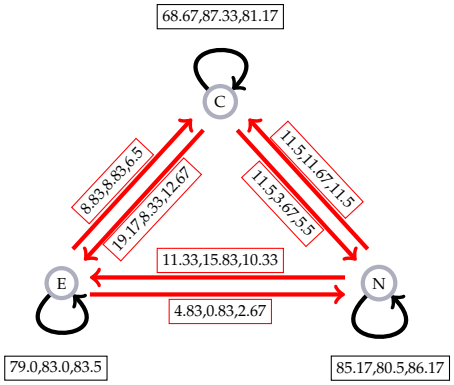


Figure 9.5: Consistency graph for predictions of ADAPET(token) versus (a) RoBERTa$_L$+CLS (b) ADAPET (CWWM) (c) ADAPET (pretrained mixNLI) in that order respectively.

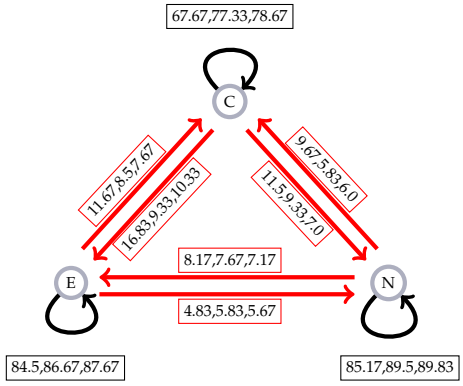


Figure 9.6: Consistency graph for predictions of ADAPET(token)+CLS versus (a) RoBERTa$_L$+CLS (b) ADAPET (CWWM)+CLS (c) ADAPET (pretrained mixNLI)+CLS in that order respectively.

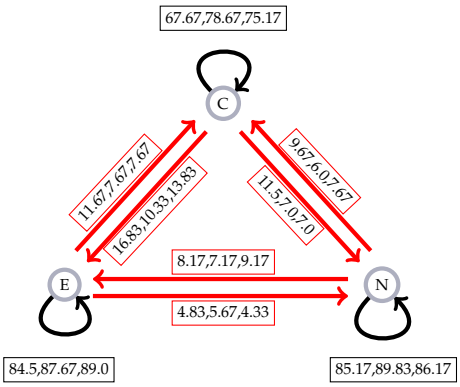


Figure 9.7: Consistency graph for predictions of ADAPET(token)+CLS versus (a) RoBERTa$_L$+CLS (b) ADAPET (pretrained mixNLI)+CLS (c) ADAPET (pretrained MNLI)+CLS in that order respectively.

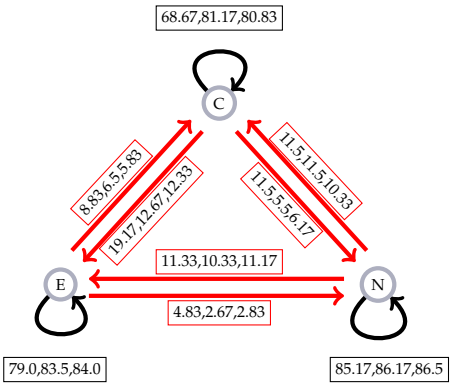


Figure 9.8: Consistency graph for predictions of ADAPET(token) versus (a) RoBERTa$_L$+CLS (b) ADAPET (pretrained mixNLI) (c) ADAPET (pretrained MNLI) in that order respectively.

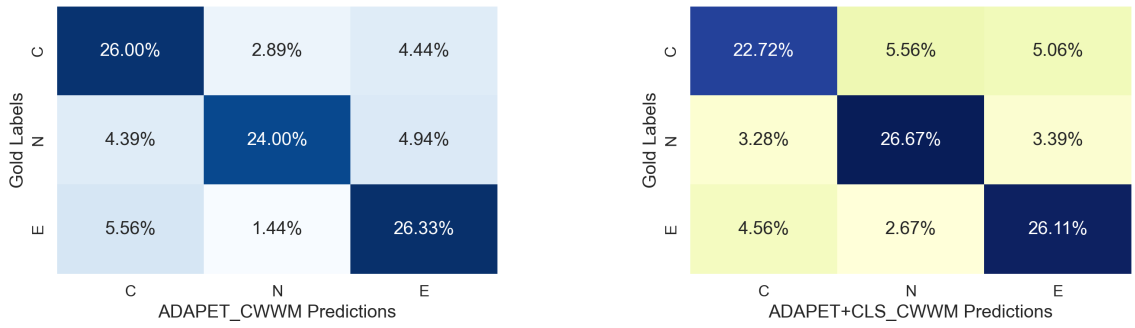


Figure 9.9: Confusion Matrix: Gold Labels versus predictions of ADAPET(CWWM), ADAPET(CWWM)+CLS.

Table 9.1: An example of tabular premise from INFOTABS [84]. The hypotheses **H1, H4** is entailed, **H2, H5** is a neutral and **H3, H6** is a contradiction. Here, the **bold** entries, which correspond to the first column, are the keys, while the corresponding entries in the second column of the same row are their respective values.

Breakfast in America	
Released	29 March 1979
Recorded	May–December 1978
Studio	The Village Recorder in LA
Genre	Pop, art rock, soft rock
Length	46:06
Label	A&M
Producer	Peter Henderson, Supertramp

H1: Breakfast in America is a pop album with a duration less than 50 minutes.
H2: Peter Henderson produces only rock albums.
H3: Breakfast in America was released towards the end of 1979.
H4: Breakfast in America is recorded in California.
H5: Supertramp is an English band.
H6: The album was released on 29 March 1978.

Table 9.2: Examples of various perturbations used to generate the adversarial test sets based on Table 9.1.

Perturb	Original text	Perturbed text
Char	Peter Henderson produces only rock albums	Peter Henbgderson produces only rock albsums Peter Hendersno produces only rokc albums Pter Henderson produces onl rock abus Petqr Henkerson prgduces only rock alocms
Loc.	Breakfast in America is recorded in California Breakfast in America is recorded in USA Breakfast in America is by an English rock band.	Breakfast in America is recorded in Florida. Breakfast in America is recorded in Syria. Breakfast in America is by an Mexican rock band.
Name	Peter Henderson produces only rock albums	John Doe produces only rock albums
Num.	The album was released on 29 March 1978.	The album was released on 29 March 346. The album was released on 1 March 1978.
Neg.	The genres of the album are pop and rock.	The genres of the album are not pop and rock.
Para	The album was recorded in the last half of 1979.	In the 2nd part of 1979, the album was recorded.

Table 9.3: More examples of various perturbations used to generate the adversarial test sets based on Table 9.1.

Perturb	Original text	Perturbed text
neg+char	The genres of the album are pop and rock.	The gejnres of the alzum are not pbp and rock.
neg+name	Peter Henderson's album was recorded in 1979.	John Doe's album was not recorded in 1979.
num+char	The album was recorded in 1979.	The album was recqorded in the last hplf of 459.
num+name	Peter Henderson's album was recorded in 1979.	John Doe's album was recorded in 731.
num+neg	The album was released on 29 March 1978.	The album was not released on 29 March 346.
num+para	The album was recorded in 1979.	In the 2nd part of 1278, the album was recorded.
para+name	Peter Henderson produces only rock albums.	Only rock albums are produced by John Doe.
num+para+name	Peter Henderson's album was recorded in 1979.	The album by John Doe was recorded in 3147.

Table 9.4: Number of examples for each perturbation type in the adversarial set.

Peturb Type	Size	Peturb Type	Size
character	1800	negation+char	1726
location	1229	negation+name	1677
name	1646	number+char	837
negation	1726	number+name	776
number	837	number+negation	817
paraphrase	1800	num+paraphrase	837
num+para+name	776	paraphrase+name	1721

Table 9.5: Top 1 accuracy of factual and relational knowledge evaluation on DRR@4.(w/o - no CLS, RoBERTa$_L$+CLS

Type	Input	RoBERTa$_L$		ADAPET	
Top 1 Accuracy		w/o	+CLS	w/o	+CLS
Factual	only E	35.5	26.2	34.3	29.2
	prem + E	59.4	29	59.7	44.8
	only C	37.2	24.6	36.9	29.8
	prem + C	54.6	26.5	49.7	39.9
	only E∪C	36.3	25.4	35.5	29.5
	prem + E∪C	57.7	27.8	54.6	42.5
Relational	only E	48.9	27	52.8	35.6
	prem + E	57.7	22.4	58.7	41
	only C	44.7	27.3	47.3	35.6
	prem + C	51.8	24	52.9	34
	only E∪C	46.7	27.2	49.9	35.6
	prem + E∪C	54.6	23.2	55.7	37.3

Table 9.6: Top 5 accuracy of factual and relational knowledge evaluation on DRR@4. (w/o - no CLS, RoBERTa$_L$+CLS.

Type	Input	RoBERTa$_L$		ADAPET	
Top 5 Accuracy		w/o	+CLS	w/o	+CLS
Factual	only E	50.4	40.6	52.4	46.6
	prem + E	72	45.3	71.5	60.7
	only C	55.2	37.4	56	47.8
	prem + C	74.6	39.3	70.2	56
	only E∪C	52.7	39.1	54.1	47.2
	prem + E∪C	73.3	42.5	70.9	58.5
Relational	only E	64.9	51.6	67.3	57.5
	prem + E	70.8	49.1	72.2	66.3
	only C	64.7	53.1	65.8	57.8
	prem + C	71.1	53.3	72	62
	only E∪C	64.8	52.4	66.5	57.6
	prem + E∪C	70.9	51.3	72.1	64.1

Table 9.7: Reasoning results on INFOTABS comparing RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training). token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking.

Splits	Premise	RoBERTa$_L$	ADAPET				ADAPET+CLS			
		+CLS	token	CWWM	+mixNLI	+MNLI	token	CWWM	+mixNLI	+MNLI
Dev	BPR	76.83	77.5	77.67	79.07	78.07	77.66	77.27	**79.63**	78.46
	DRR@4	76.39	76.67	76.97	78.57	77.33	76.88	77.11	**78.64**	77.44
	DRR@8	75.36	77.77	77.63	78.83	77.93	77.81	77.57	**79.42**	78.96
α_1	BPR	75.29	76.87	75.93	77.33	77.47	77.47	78.05	77.96	**78.33**
	DRR@4	75.78	77.5	77.53	**78.6**	78.17	77.18	77.66	78.04	78.13
	DRR@8	75.61	78.3	78	79	78.2	78.03	78.7	78.63	**79.05**
α_2	BPR	66.5	67.93	68.07	**72.4**	69.8	68.48	69.55	72.16	70.09
	DRR@4	67.22	69.33	69	70.23	69.03	68.92	68.29	**70.58**	69.24
	DRR@8	67.11	69.43	69.37	71.87	69.97	69.24	69.81	**72.13**	70.61
α_3	BPR	64.26	63.73	64.6	66.23	64.13	64.98	65.67	**68.4**	66.03
	DRR@4	64.88	67.43	67.5	68.7	67.33	66.02	66	**68.74**	67.37
	DRR@8	67.53	68.07	67.63	**70.2**	68	66.66	67.59	69.2	68.31

Table 9.8: Results on linearized table comparing [84] and our approach (ADAPET).

Test Splits	[84]	Ours
Dev	**77.61**	76.7
α_1	75.06	**76.1**
α_2	69.02	**69.6**
α_3	64.61	**65.3**

Table 9.9: Adversarial Reasoning results on perturbed sets with DRR@8 comparing RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training), token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking. Rows in the tables are sorted in ascending order w.r.t RoBERTa$_L$+CLS performance.

Perturb	RoBERTa$_L$	ADAPET				ADAPET+CLS			
	+CLS	token	CWWM	+mixNLI	+MNLI	token	CWWM	+mixNLI	+MNLI
num+para+name	13.04	10.1	7.1	11.7	10.1	11.7	13.81	**16.62**	13.55
number+name	15.72	14.6	9.0	14	13.2	15.6	15.36	**18.94**	15.85
negation+name	19.08	16.1	7.2	**20**	11.6	14.43	12.88	14.37	12.1
num+paraphrase	27.46	59.5	**61.0**	58.4	57.3	52.5	51.49	56.63	54.95
paraphrase+name	30.79	22.6	18.3	28.3	24.9	27.01	27.3	**30.85**	27.71
name	32.7	24.7	19.0	31.1	28	28.9	29.96	**33.44**	30.69
random	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33
number+negation	36.13	42.7	31.8	**53.2**	28.3	37.91	47.32	37.75	24.04
negation+char	39.39	41.4	38.5	**47.6**	40.1	42.9	41.94	42.06	40.85
negation	53.7	58.1	53.3	**64.8**	56.1	57.6	56.83	59.15	53.88
number+char	54.43	58.8	**65.2**	57.1	60.3	55.79	47.9	57.1	59.28
number	56.1	57.8	**62.0**	57.8	57	52.44	51.37	55.79	54.6
character	63.05	62.8	63.3	65.9	64.4	64.05	64.44	66.05	**66.83**
location	67.6	70	**70.2**	67.7	69.1	69.81	66.8	67.4	65.98
paraphrase	70.56	72.3	73.2	**73.8**	73.4	71.6	70.5	72.66	72.3
INFOTABS (α_1)	76.56	78.1	78.9	**80.2**	78.9	78.27	77.66	78.5	78.66

Table 9.10: Adversarial reasoning results on perturbed sets with DRR@4 RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training). token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking. Rows in the tables are sorted in ascending order w.r.t RoBERTa$_L$+CLS performance.

Perturb	RoBERTa$_L$	ADAPET				ADAPET+CLS			
	+CLS	token	CWWM	+mixNLI	+MNLI	token	CWWM	+mixNLI	+MNLI
number+name	14.17	20	12.9	14.5	18.3	17.78	17.13	**20.8**	16.49
num+para+name	15.08	16.3	8.7	9.5	15.2	15.08	16.88	**17.9**	11.25
negation+name	18.66	17.1	13.9	7.8	11.6	**18.48**	13.23	10.31	10.55
number+negation	28.63	36.9	43.2	41.5	23.1	39.31	**45.86**	37.91	25.78
paraphrase+name	30.9	32.3	22.6	26.7	27.4	32.2	32.36	**32.48**	26.55
name	32.4	32.1	25.7	29.8	30.5	33.56	33.6	**33.7**	30.01
random	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33
negation+char	40.38	42.5	41.1	39.7	37.4	**45.4**	40.61	40.49	38.9
negation	46.46	**59.4**	57	56	52	59.03	56.89	58.4	55.7
num+paraphrase	52.56	57.3	59.5	58.4	**59.4**	57.7	51.86	51.13	48.9
number+char	53.34	55.5	63.2	61.6	**64.8**	55.3	49.81	55.85	54.9
number	54.9	59.5	59.1	56.9	**59.8**	55.91	52.09	51.97	51.13
character	56.88	63.7	63.7	**67.1**	63.3	65.16	60.88	65.16	65.27
paraphrase	66.3	72.5	72.9	**73.1**	72.2	69.88	68.44	73.1	72.22
location	69.65	**73**	71.2	70	69.9	69.97	65.825	68.59	68.1
dev	76.39	76.4	77.8	**78.2**	77.2	76.27	78.05	78.16	77.5
α_1	75.78	76.5	78	**79.4**	79.2	76.44	77.66	78.22	78.11

Table 9.11: Adversarial Reasoning results on perturbed sets with BPR comparing RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training). token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking. Rows in the tables are sorted in ascending order w.r.t RoBERTa$_L$+CLS performance.

Perturb	RoBERTa$_L$	ADAPET				ADAPET+CLS			
	+CLS	token	CWWM	+mixNLI	+MNLI	token	CWWM	+mixNLI	+MNLI
negation+name	11.74	10.4	10.2	**21.1**	15.6	17.35	14.37	13.89	12.93
num+para+name	14.06	10.6	8.4	**20.7**	12	17.13	16.88	14.83	13.04
number+name	17.26	12.5	10.2	**20.9**	14.8	18.42	18.81	18.42	16.88
paraphrase+name	33	25.8	20.6	**37.6**	31.5	31.2	33.41	32.1	31.3
random	33.33	33.33	33.3	33.33	33.33	33.33	33.33	33.33	33.33
name	34.6	26.5	20.4	**36.4**	33.4	32.41	34.82	33.96	33.2
negation+char	37.71	38.5	40.3	**47.8**	41.3	43.56	40.21	41.25	40.49
number+negation	38.36	30.2	48.7	**54.8**	30.1	37.69	47.26	38.7	26.06
negation	48.9	54.2	57.2	**65.4**	55.3	58.27	55.27	58.45	55.6
number	56.63	**62.3**	55.8	51.9	56	55.43	50.53	53.52	56.1
num+paraphrase	56.98	**62.3**	57.6	49.7	54.5	55.55	49.34	52.26	55.19
number+char	59.11	**66.1**	60.3	45.1	55.6	55.9	49.32	52.46	60.2
character	61.5	64.1	62.5	64.4	66.1	64.9	63.16	**66.61**	65.94
location	68.2	72.4	**72.7**	68.1	70.1	69.08	67.69	66.47	69.48
paraphrase	68.44	72.3	71.8	**72.6**	72.3	72.05	70.33	71.7	**72.66**
dev	76.83	78.1	76.4	**79.8**	79.1	78.72	78.05	79.22	78.55
α_1	75.29	78.1	76.1	77.4	77.4	77.38	77.83	78	**78.38**

Table 9.12: Results on Label Correctness (% of our generated labels match with human's predictions) and average Grammar score (out of 5) from human evaluation.

Perturbation	Label Correctness(%)	Grammar Score
character	99	4.46
location	79	4.5
name	97	4.5
negation	93	4.36
number	81	4.5
paraphrase	89	4.42
negation+char	88	4.3
negation+name	96	4.5
number+char	77	4.3
number+name	96	4.5
number+negation	80	4.44
num+paraphrase	77	4.48
num+para+name	95	4.42
paraphrase+name	94	4.5

Table 9.13: Results on two label classification (Entailment & Contradiction).

Splits	RoBERTa$_L$+CLS	ADAPET		
	DRR@4	BPR	DRR@4	DRR@8
Dev	81.5	83.5	**84.3**	82.8
α_1	80.25	83.8	**84.3**	**84.3**
α_2	64.66	65.9	66.9	**67.7**
α_3	76	75.1	**78.5**	77.4

Table 9.14: Reasoning wise number of correct predictions of DRR@4 on subset of dev set, (a, b, c) are human prediction count.

Reasoning Type	ENTAILMENT					NEUTRAL					CONTRADICTION				
	RoBERTa$_L$	ADAPET		ADAPET+CLS		RoBERTa$_L$	ADAPET		ADAPET+CLS		RoBERTa$_L$	ADAPET		ADAPET+CLS	
	+CLS	token+mixNLI		token+mixNLI		+CLS	token+mixNLI		token+mixNLI		+CLS	token+mixNLI		token+mixNLI	
Numerical (11, 3, 7)	9	9	10	10	8	3	2	3	3	3	6	6	4	6	5
Lexical Reasoning (5, 3, 4)	5	4	4	3	5	2	1	1	1	2	2	3	2	3	3
Subjective/OOT (6, 41, 6)	3	3	3	3	3	37	36	36	37	35	4	4	1	3	5
KCS (31, 21, 24)	25	21	26	20	25	20	20	18	19	18	21	22	18	21	21
Temporal (19, 11, 25)	16	13	15	15	14	7	6	5	6	7	18	20	15	17	17
Multirow (20, 16, 17)	13	12	15	15	13	13	12	11	11	13	15	16	14	15	13
Coref (8, 22, 13)	5	6	5	6	6	19	20	18	20	18	7	10	8	7	8
Quantification (4, 13, 6)	2	2	2	2	2	11	11	12	12	12	2	3	3	3	3
Named Entity (2, 2, 1)	1	2	2	1	2	1	1	2	1	1	1	1	1	1	1
Simple Lookup (3, 0, 1)	2	3	3	2	3	0	0	0	0	0	0	0	0	0	0
Negation (0, 0, 6)	0	0	0	0	0	0	0	0	0	0	4	6	5	5	4
Entity Type (6, 8, 6)	6	5	5	4	6	7	7	7	7	7	6	6	5	6	4

Table 9.15: Category wise accuracy scores of DRR@4 on dev set.

Categories	ENTAILMENT					NEUTRAL					CONTRADICTION				
	RoBERTa$_L$	ADAPET		ADAPET+CLS		RoBERTa$_L$	ADAPET		ADAPET+CLS		RoBERTa$_L$	ADAPET		ADAPET+CLS	
	+CLS	token	+mixNLI	token	+mixNLI	+CLS	token	+mixNLI	token	+mixNLI	+CLS	token	+mixNLI	token	+mixNLI
Album	71	79	74	76	81	76	86	88	86	93	60	79	79	74	74
Animal	78	81	89	89	85	70	81	81	85	81	56	70	74	81	78
City	59	63	63	57	69	67	80	65	71	75	53	61	63	65	55
Country	78	75	83	64	78	56	67	64	61	72	56	69	72	58	67
Food&Drinks	96	88	88	88	88	67	75	75	71	79	83	88	79	71	71
Movie	85	75	83	80	80	75	85	70	82	73	62	75	80	73	80
Musician	87	78	84	83	88	86	90	85	89	89	75	83	79	78	78
Organization	83	50	100	75	92	58	75	50	83	75	58	58	58	50	50
Painting	78	81	81	81	85	93	93	93	96	93	78	89	85	78	85
Person	74	73	78	74	78	81	85	80	78	81	67	79	76	77	74
Others	71	69	82	69	80	64	78	69	73	73	49	73	69	67	60

CHAPTER 10

XINFOTABS: MULTILINGUAL TABULAR INFERENCE

Adapted from B. Minhas, A. Shankhdhar, A. Gupta, D. Aggarwal, and S. Zhang, *XInfoTabS: Evaluating multilingual tabular natural language inference*, in Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER), Dublin, Ireland, May 24, 2022, Association for Computational Linguistics, pp. 59–77.

The recent development of multi-lingual extensions of contextualizing models such as mBERT [51] from BERT and XLM-RoBERTa [40] from RoBERTa, has led to substantial interest in the problem of multi-lingual NLI and the creation of multi-lingual XNLI [42] and TaxiXNLI [117] dataset from English MNLI [280] dataset. However, there is still no equivalent multi-lingual NLI dataset for semi-structured tabular data. To fill this gap, we propose XINFOTABS, a multi-lingual extension of INFOTABS dataset. The XINFOTABS dataset consists of ten languages, namely English ('en'), German ('de'), French ('fr'), Spanish ('es'), Afrikaans ('af'), Russian ('ru'), Chinese ('zh'), Korean ('ko'), Hindi ('hi') and Arabic ('ar'), which belong to seven distinct language families and six unique writing scripts. Furthermore, these languages are the majority spoken in all seven continents covering 2.76 billion native speakers in comparison to 360 million English language (INFOTABS) speakers. Refer to Table 10.1 for more information.

The intuitive method of constructing XINFOTABS, i.e., human-driven manual translation, is too expensive in terms of money and time. Alternatively, various state-of-the-art machine translation models, such as mBART50 [251], MarianMT [116], M2M100 [60], have greatly enhanced translation quality across a broad variety of languages. Furthermore, NLI requires simply that the translation models retain the semantics of the premises and hypotheses, which machine translation can deliver [117]. Therefore, we use automatic machine translation models to construct XINFOTABS from INFOTABS.

Tabular data is far more challenging to translate than semantically complete and grammatical sentences with existing state-of-the-art translation systems. To mitigate this challenge, we propose an efficient, high-quality translation pipeline that utilizes Name Entity Recognition (NER) and table context in the form of category information to convert table cells into structured sentences before translation. We assess the translations via several automatic and human verification methods to ensure quality. Our translations were found to be accurate for the majority of languages, with German and Arabic having the most and least exact translations, respectively. Table 10.2 shows an example from the XINFOTABS dataset.

We conduct tabular NLI experiments using XINFOTABS in monolingual and multilingual settings. By doing so, we aim to assess the capacity and cross-lingual transferability of state-of-the-art multilingual models such as mBERT [51], and XLM-Roberta [40]. Our investigations reveal that these multilingual models, when assessed for additional languages, perform comparably to English. Second, the translation-based technique outperforms all other approaches on the adversarial evaluation sets for multilingual tabular NLI in terms of performance. Thirdly, the method of intermediate-task finetuning, also known as pre-finetuning, significantly improves performance by finetuning on additional languages prior to the target language. Finally, these models perform admirably on cross-lingual tabular NLI (tables and hypotheses given in different languages), although the additional effort is required to improve them.

10.1 Contributions

We make the following contributions through this chapter [1]:

1. We introduce XINFOTABS, a multi-lingual extension of INFOTABS, a semi-structured tabular inference English dataset over ten diverse languages.
2. We propose an efficient pipeline for high-quality translations of semi-structured tabular data using state-of-the-art translation models.
3. We conduct intensive inference experiments on XINFOTABS and evaluate the performance of state-of-the-art multilingual models with various strategies.

[1]The dataset and associated scripts, is available at https://xinfotabs.github.io/.

This work is published at FEVER 2022 workshop at ACL 2022 as [171]. We also construct EI-InfoTabS: an English-Indic bilingual tabular natural language inference dataset (TNLI), in which the tabular premise is in English language and hypothesis Indic languages. To create EI-InfoTabS we translate the textual hypotheses of the English TNLI dataset InfoTabS into eleven major Indian languages. This work was published at NAACL 2022 as [4].

10.2 Background

Given the need for greater inclusivity towards linguistic diversity in NLP applications, various multilingual versions of datasets have been created for text classification [42, 203, 290], question answering [8, 37, 138] and structure prediction [188, 211]. Following the introduction of datasets, multilingual leaderboards like XTREME leaderboard [98], the XGLUE leaderboard [147] and the XTREME-R leaderboard [228] have been created to test models' cross-lingual transfer and language understanding.

Multilingual models can be broadly classified into two variants: (a) Natural Language Understanding (NLU) models like mBERT [51], XLM [41], XLM-R [40], XLM-E [32], RemBERT [35], and (b) Natural Language Generation (NLG) models like mT5 [286], mBART [157], M2M100 [59]. NLU models have been used in multilingual language understanding tasks like sentiment analysis, semantic similarity and natural language inference while NLG models are used in generation tasks like question-answering and machine translation.

10.2.1 Machine Translation

Modern machine translation models involve having an encoder-decoder generator model trained on either bilingual [258] or a multilingual parallel corpus with monolingual pretraining e.g. mBART [157] and M2M100 [59]. These models have been shown to work very well even for low-resource languages due to cross-language transfer properties. Recently auxiliary pertaining for machine translation models have garnered attention, with a focus on autonomous quality estimation metrics [64, 241, 242]. As such, automatic scores like the BERTScore [304], BLEURT [233] and COMET Score [216] have high human evaluation correlation, are increasingly used to assess NLG tasks.

10.3 Why the INFOTABS Dataset?

There are only two public datasets, both in English, available for semi-structured tabular reasoning, namely TabFact [26] and INFOTABS [84]. We choose INFOTABS because it includes multiple adversarial test sets for model evaluation. Additionally, the INFOTABS dataset also includes the NEUTRAL label, which is absent in TabFact. The INFOTABS dataset contains 2,540 tables serving as premise and 23,738 hypothesis sentences along with associated inference labels. The table-sentence pairs are divided into development, and three evaluation sets α_1, α_2, and α_3, each containing 200 unique tables along with nine hypothesis sentences equally distributed among three inference labels (ENTAIL, CONTRADICT, and NEUTRAL). α_1 is a conventional evaluation set that is lexically similar to the training data. α_2 has lexically adversarial hypotheses. And α_3 contains domain topics that are not present in the training set. The remaining 1,740 tables with corresponding 16,538 hypotheses serve as a training set. Table 10.3 describes the inference performance of RoBERTa$_L$ model on INFOTABS dataset. As we can see, the Human Scores are superior to that of RoBERTa$_L$ model trained with TabFact representation. Since the XINFOTABS is translated directly from the INFOTABS, we expect a similar human baseline for XINFOTABS.

10.4 Table Representation

Machine translation of tabular data is a challenging task. Tabular data is semi-structured, non-sentential (ungrammatical), and succinct. The tight form of tabular cells provides inadequate context for today's machine translation models, which are primarily designed to handle sentences. Thus, table translation requires additional context and conversion. Furthermore, frequently occurring named entities in tables must be transliterated rather than translated. Figure 10.1 shows the table translation pipeline. We describe our approach to context addition and handling of named entities in detail in the following subsections §10.4.1.

10.4.1 Table Translation Context

There are several ways to represent tables, each with its own set of pros and cons, as detailed below:

10.4.1.1 Without context

The most straightforward way to represent a table would be to treat every key (header) and value (cell) as separate entities and then translate them independently. This approach results in poor translations as the models have no context regarding the keys. The key *"Length"* in English in context of *Movies* would correspond to *"durée"*, meaning *duration* in French but in *Object* context, would correspond to *"longueur"*, meaning *size or span*. Thus, context is essential for accurate table translation.

10.4.1.2 Full table

Before transferring data from the header and table cells to translation models, one may concentrate and seam each table row using a delimiter such as a colon (":") to separate key from value and a semi-colon (";") to separate rows [26]. This method provides full context and completely translates all table cells. However, in practice, this strategy has two major problems:

a. *Length Constraint*: All transformer-based models have a maximum input string length of 512 tokens.[2] Larger tables with tens of rows may not be translated using this approach.[3] In practice, strings longer than 256 tokens have been shown to have inferior translation quality.[4]

b. *Structural Issue*: When a linearized table is directly translated, the delimiter tokens (":" and ";") get randomly shifted.[5] The delimiter counts are also altered. Hence, the translation appears to merge characters from adjacent rows, resulting in inseparable translations. Ideally, the key and value delimiter token locations should be invariant in a successful translation.

---

[2]Recently, models bigger than 512 tokens have been developed, e.g. [10, 13], but no publicly accessible long-sequence ($>$ 512 tokens) multilingual machine translation model exists at the moment.

[3]Average # of rows in InfoTabS is: 8.8 for Train, Development, α_1 and α_2, and 13.1 for α_3.

[4] [182] raises a similar issue for NLI.

[5]Using "—" instead of ":" helps key-value separation.

10.4.1.3 Category context

Given the shortcomings of the previous two methods, we devise a new strategy: we add a *general context* that describes table rows at a high level to each linearized row cell. We leverage the *table category* here, as it offers enough context to grasp the key's meaning. For the key *"Focus"* in Table 10.2, the category information *Sports* offers enough context to understand its significance in relation to boxing. The context added representation for this key-value pair will be *"Sports — Focus — Punching , Striking"*. We use "—" delimiter for separating the context, key, and value. Furthermore, multiple values are seperated by ",". Unlike full table translation, row structure is preserved since each row is translated independently and no row surpasses the maximum token limit. We observe an average increase of 5.5% in translation performance (cf. §10.5).

10.4.2 Handling Named Entities

Commercial translation methods, like Google Translate, correctly transliterate specified entities (such as proper nouns and dates). However, modern open-source models like mBART50 and M2M100 translate name entity labels, lowering overall translation quality. For example, *Alice Sheets* is translated to *Alice draps* in French. We propose a simple preprocessing technique to address the transliterate/translate ambiguity. First, we use the Named Entity Recognition (NER) model[6] [110] to identify entity information that must be transliterated, such as proper nouns and dates. Then, we add a unique identifier in the form of double quotations (" "), e.g., *"Alice Sheets"*, and apply the translation model. Finally, we delete the quotation mark (" ") from the translated sentence after it has been translated. This helps the models identify these entities easily due to their pre-training.

10.5 Translation and Verification

As mentioned previously, we now grasp how to represent a table. Consequently, these reformatted tables can now be fed into reliable translation models. To accomplish this, we assess many prominent multilingual (e.g., mBART50 [251] and M2M100 [61]) and bilingual (e.g., MarianMT [116]) translation models as described below:

[6] spaCy NER tagger

10.5.1 Multilingual Models

This category of models used includes widely used machine translation models trained on a large number of languages such as mBART50 [251] which can perform translation between any two languages from the list of 50 languages and M2M100 [61] which has 100 training languages. Apart from these models, we used Google Translate[7] to compare against our dataset translation quality.

10.5.2 Bilingual Models

Earlier studies have revealed that bilingual models outperform multilingual models in machine translation of high-resource languages. Thus, for our experiments, we also considered language-specific bilingual translation models in MarianMT [116] repository. Because the MarianMT models were not available for a few languages (e.g., Korean (ko)) of XINFOTABS, we could not conduct experiments for some languages.

We also use an efficient data sampling technique to determine the ideal translation model for each language, as detailed in the next section. The results for the translations are shown in Table 10.4.

10.5.3 Translation Model Selection

Translating the complete INFOTABS dataset to find the optimal model is practically infeasible. Thus, we select a representative subset of the dataset that approximates the full dataset rather well. Finally, we use optimal models to translate the complete INFOTABS dataset. The method used for making the subset is discussed in the ***Table Subset Sampling Strategy*** and ***Hypothesis Subset Sampling Strategy*** sections given below:-

10.5.3.1 Table subset sampling strategy

In a table, keys can serve as an excellent depiction of the type of data included therein. For example, if the key *"children"* is used, the associated value is almost always a valid *Noun Phrase* or a collection of them. Additionally, the type of keys for a given category remains constant across tables, but the values are always different.[8] This fact is used to

[7] https://translate.google.co.in/

[8] There are 2,163 unique keys in INFOTABS.

sample a subset of diverse tables based on keys and categories. Specifically, we sample tables for each category based on the frequency of occurrence of keys in the dataset to guarantee diversity. The sum of the frequencies of all the keys in a table is computed for each table. Finally, the top 10% of tables with the largest frequency sum in each category are chosen to be included in the subset. In the end, we construct a subset with 11.14% tables yet containing 90.2% of the all unique keys.

10.5.3.2 Hypothesis subset sampling strategy

To get a diverse subset of hypotheses, we employ Top2Vec [7] embedding for each hypothesis, then use k-means clustering [112] to choose 10% of each cluster. Sampling from each cluster ensures we cover all topics discussed in the hypothesis, resulting in a subset of 2,569 hypothesis texts.

10.5.3.3 Model selection strategy

To choose the translation model that will be used to generate the language datasets, we first translate the premise and hypothesis subsets for all languages using each of the existing models, as described before. Following translation, we compute the various scores detailed in Section 10.5.4. Finally, the model with the highest average of premise and hypothesis translation ***Human Evaluation Score*** for the specified language is chosen to translate the complete INFOTABS datasets.

10.5.4 Translation Quality Verification

With the emergence of Transformer-based pre-trained models, significant progress has been made in automated quality assessment using semantic similarity and human sense correlation [19] for machine translation evaluation. To verify our created dataset XINFOTABS, we use three automated metrics in addition to human ratings.

10.5.4.1 Paraphrase score (PS)

PS indicates the amount of information retained from the translated text. To capture this, we estimate the cosine similarity between the original INFOTABS text and the back-translated English XINFOTABS text sentence encodings. We utilize the all-mpnet-v2 [239] model trained using SBERT [23] method for sentence encoding.

10.5.4.2 Multilingual paraphrase score (mPS)

Different from PS, mPS directly uses the multilingual XINFOTABS text instead of the English back-translated text to compare with INFOTABS text. We produce sentence encodings for multilingual semantic similarity using the multilingual-mpnet-base-v2 model [86] trained using the SBERT method.

10.5.4.3 BERTScore (BS)

BERTScore is an automatic score that shows high human correlation and has been a widely used quality estimation metric for machine translation tasks [304].

10.5.4.4 Human evaluation score (HES)

We hired five annotators to label sampled subsets of 500 examples per model and language. Human verification is accomplished by supplying sentence pairs and requesting that annotators classify them as identical or dissimilar based on the meaning expressed by the sentences.

10.6 Human Annotation Guidelines

We employed five undergraduate students proficient in English as human evaluation annotators. They were presented with an instruction set with sample examples and annotations before the actual work. We paid the equivalent of 10 cents for every labeled example. The study's authors reviewed random annotations to confirm their quality.

Annotation Guidelines: We refer to the work by [126] while setting up our annotation task and instruction guidelines. We gathered 500 table-sentence pairs representing original (en) and back-translated (en) texts per model-language into several Google spreadsheets. We had a total of 108 sheets (4 models, 9 languages, 3 Modes (table-keys, table-values, and hypothesis) and hence 54000 annotation instances. Each sheet was assigned to a single annotator, who was required to adhere to the semantic similarity task requirements, which are outlined below:

1. The Semantic Similarity task requires the annotator to classify each sentence-pair as conveying the same meaning (label 1) or conveying different meaning (label 0) than each other.

2. In case their exists a difference of syntax including spelling mistakes, punctuation

error or missing special characters, the annotators was asked to ignore these as long as the sentence meaning is understandable (label 1). In case proper nouns were misspelled, the annotator must judge the spellings as phonetically similar (label 1) or not (otherwise label 0).

3. The annotators were asked to be lenient on the grammar, allowing for active-passive changes and tense change, if the sentences convey close to the same meaning i.e. (label 1).

4. In case acronyms or abbreviations were present in the sentences, the annotators were asked to mark them as same (label 1) if the sentences had proper expansion/contractions.

5. In presence of numbers or dates, the annotators were asked to be extremely strict and label even slightly differing dates or numbers like (XXXI v.s. 30) as completely different (label 0).

6. In case of any further ambiguity, the judgement was left to the annotators human far-sight as long as the adhere to the task definition.

We estimated the accuracy of human verification for every models and languages by averaging the annotator labels.

Analysis: We arrive at an average language score of 85 for tables and 91 for hypotheses for the final selected models in all languages. The results are summarised in Table 10.4. These results are also utilized to determine the optimal models for translating the entire dataset. MarianMT is used to create the entire dataset in German, French, and Spanish, mBART50 is used to create the Tables dataset in Afrikaans, Korean, Hindi, and Arabic, and M2M100 is used to create the entire dataset in Russian and Chinese, as well as the hypothesis dataset in Afrikaans, Korean, Hindi, and Arabic.

10.7 Experiment and Analysis

In this section, we study the task of Multilingual Tabular NLI, utilizing our XINFOTABS dataset as the benchmark for a variety of multilingual models with multiple training-testing strategies. By doing so, we aim to assess the capacity and cross-lingual transferability of state-of-the-art multilingual models. For the inference task, we linearize the table using the *"Table as Struct"- TabFact* described in INFOTABS.

10.7.1 Multilingual Models

We use pre-trained multilingual models for all our inference label prediction experiments. We use a multilingual mBERT-base (cased) [51] model pre-trained on masked language modeling. This model will be referred to as $\text{mBERT}_{\text{BASE}}$. The other model we evaluated is the XLM-RoBERTa Large (XNLI) model [40], which is trained on masked language modeling and then finetuned for the NLI task using the XNLI dataset. This model is referred to as XLM-R Large (XNLI).

Hyperparameter: The $\text{XLM-R}_{\text{LARGE}}$ (XNLI) model was taken from HuggingFace[9] models and finetuned using PyTorch Framework[10] on Google Colaboratory[11] which offer a single P100 GPU. We utilized accuracy as our metric of choice, same as . We used Adagrad [144] as our optimizer with a learning rate of $1 * 10^{-4}$. We ran our finetuning script for ten epochs with a validation interval of 1 epoch, and early stopping callback enabled with the patience of 2. Given the large model size, we had to use a batch size of 4.

The $\text{mBERT}_{\text{BASE}}$ (cased) model was trained on TPUv2 8 cores using the PyTorch Lightning[12] Framework. AdamW [163] was our choice of optimizer with learning rate $5 * 10^{-6}$. We ran our finetuning script for ten epochs with a validation interval of 0.5 epochs, and early stopping callback enabled with the patience of 3. Given the model's small size, we used a batch size of 64 (8 per TPU core).

Tables 10.5, 10.6, and 10.7 show the performance of the discussed multilingual models for α_1, α_2, and α_3 test splits respectively. Tables 10.6 and 10.7 show the results for all baseline tasks on the Adversarial Validation Sets α_2 and α_3. On all three evaluation sets, regardless of task type, the $\text{XLM-RoBERTa}_{\text{Large}}$ model outperforms mBERT. This might be because XLM-RoBERTa has more parameters, and is better pre-trained and pre-tuned for the NLI task using the XNLI dataset.

---

[9] huggingface.co

[10] pytorch.org

[11] Google Colaboratory

[12] PyTorch Lightning

10.7.2 Using English Translated Test Sets

We aim to investigate the following question: *How would models trained on original English* INFOTABS *perform on English translated multilingual* XINFOTABS*?*. We trained multilingual models using the original English INFOTABS training set, and used the English translated XINFOTABS development set, and three test sets during the evaluation. According to Table 10.5, German has the best language-wise performance for α_1. From Table 10.6, German, French, and Afrikaans have the highest average scores for α_2. French and Russian have the best scores on α_3 as shown in Table 10.7. Arabic has the lowest average of any language across all three test sets. Here, the model trained on English INFOTABS is being used for all the languages. Since the model is the same for all languages, the variation in performance only depends on English translation across XINFOTABS languages. On α_2 and α_3 sets, this task on average performs competitively against all other baseline tasks.

10.7.3 Language-Specific Model Training

In this subsection, we try to answer the question: *Is it beneficial to train a language-specific model on* XINFOTABS*?* In doing so, we finetune ten distinct models, one for each language on XINFOTABS. Comparing models on this task helps comprehend the model's intrinsic multilingual capabilities for tabular reasoning. Among the language-specific models, English has the best language average in all three test sets, while Arabic has the lowest.

Additionally, there is a substantial variation in the quality of translation and model multilingualism competence. The high-resource languages often perform better since the pre-trained models have been trained on a larger amount of data from these languages. Surprisingly, §10.7.3 setting has lower average mBERT scores for all three splits than §10.7.2 setting. The benefit of training the model in English seems to surpass any loss incurred during translating test sets into English. However, this is not the case with XLM-R(XNLI). The average scores increase substantially for α_1 split in §10.7.3 setting compared to §10.7.2 setting, decrease slightly for α_2, and remain constant for α_3. The α_1 set improves due to its similar split to the train set, whereas the α_2 set slightly worsens since it includes human-annotated perturbed hypotheses with labels flipped. Lastly, the α_3 set comprises tables from zero-shot domains i.e. unseen domain tables, so it remains constant. Our exploration of models' cross-lingual transferability is provided in Appendix F.

10.7.4 Fine-Tuning on Multiple Languages

Earlier findings indicate that fine-tuning multilingual models for the same task across languages improves performance in the target language [201, 205, 268]. Thus, *do models benefit from sequential fine-tuning over several* XINFOTABS *languages?* To answer it, we investigate this strategy of pre-finetuning in two ways, (a) by using English as the predominant language for pre-finetuning, and (b) by utilizing all XINFOTABS languages to train a unified model, .

10.7.4.1 Using English language

We fine-tune our models on the English INFOTABS and then on XINFOTABS in each language individually. Thus, we train nine models in total, one for each multilingual language (except English). English was chosen as the pre-finetuning language due to its strong performance in the §10.7.3 paradigm and prior research demonstrating English's superior cross-lingual transfer capacity [201]. Across all three splits, the average score improves from the §10.7.3 setting, demonstrating that pre-finetuning the English dataset benefits other multilingual languages. The most significant gains are shown in lower resource languages, notably Arabic, which improved by 3% for α_1, 2% for α_2, and 1% for α_3 in comparison to the §10.7.3 approach.

10.7.4.2 Unified model approach

We explore whether fine-tuning on other languages is beneficial, where we fine-tune a single unified model across all XINFOTABS languages' training sets and use it for making predictions on XINFOTABS test sets. We observe that the finetuning language order affects the final model performance if done sequentially. We find that training from a high to a low resource language leads to the highest average accuracy improvement. This is due to the catastrophic forgetting trait [80], which encourages training on more straightforward examples first, i.e., those with better performance. Hence, we trained in the following language order: en $\rightarrow$ fr $\rightarrow$ de $\rightarrow$ es $\rightarrow$ af $\rightarrow$ ru $\rightarrow$ zh $\rightarrow$ hi$\rightarrow$ ko $\rightarrow$ ar.

We observe that the XLM RoBERTa Large model performs the best across all baseline tasks in the α_1 set. On average, this performance is comparable to English pre-finetuning. While the accuracy of high resource languages remains constant or marginally declines compared to the §10.7.3 setting, there is a substantial improvement in accuracy for low

resource languages, particularly Arabic, which increases by 2%. It performs similarly to English pre-finetuning. To conclude, more fine-tuning is not always beneficial for all models, but it benefits larger models like the XLM-R Large. Models improve performance for low-resource languages compared to the §10.7.3 setting (i.e., no pre-finetuning), but not nearly as much as that of English-based pre-finetuning.

10.7.5 English Premise Multilingual Hypothesis

The premise of English's multilingual hypothesis is practical, as it is frequently observed in the real world. The majority of the world's facts and information are written in English. For instance, Wikipedia has more tables in English than in any other language, and even if a page is available, it is likely that it missing an infobox. However, because people are innately bilingual, inquiries or verification queries concerning these facts could be in a language other than English. As a result, the task of developing cross-lingual tabular NLI is critical in the real world.

To study this problem, we look at the following question: *How effective are models with premise and hypothesis stated in distinct languages?* To answer this, we train the models using the original INFOTABS premise tables in the English language and multilingual hypotheses in XINFOTABS, i.e., nine languages. We note that XLM-R Large (XNLI) has the highest accuracy for the α_1 set. On average, the high-resource languages German, French, and Spanish perform favorably across models, whereas Arabic underperforms. Both models have shallow scores in German for the α_2 set, which defy earlier observations. This might be because the adversarial modifications in the α_2 hypothesis might not be reflected in the German translation. XLM-R Large has the highest accuracy on this set, with French and Spanish being the most accurate languages. The models for the α_3 validation set demonstrate that language average accuracy is nearly proportional to the size of translation resources. However, the scores are marginally lower on average for the α_2 set.

Surprisingly, models perform worse on average than with §10.7.3 setting on the α_1 and α_2 sets while performing similarly on the α_3 set. Except for α_2 on German, the average language accuracy changes are directly proportional to the language resource, implying that the constraint could be translation quality; left for future study.

10.7.6 Robustness and Consistency

In this part, we examine the findings for several languages and delve a little more into the key disparities in performance across them. We compare the results of the experiments for §10.7.3 setting for α_1 set of best-performing language (en) with three languages - (a) A high resource language (fr), (b) A mid resource language (af) and c) A low resource language (hi). We compute four numbers for each of the languages (l) (where l is (fr), (af), or (hi)) and (en) - the proportion of instances when (a) both are right, (b) both are erroneous (c) correct (en) but incorrect (l), and (d) correct (l) but incorrect (en). We compute this number overall as well independently for each of the inference labels, refer Figure 10.2.

We note that the majority of instances were correctly categorized in both English and all three other languages. This is followed by the number of instances in which English and all other languages categorised examples inaccurately. Additionally, we notice a greater proportion of samples that are correctly identified by English but wrongly classified by all other languages, as opposed to the contrary. Furthermore, the label NEUTRAL has the highest proportion of correctly classified examples across all languages, whereas the label CONTRADICT has the lowest.

In Figure 10.3, we notice that the CONTRADICT gets confused a lot with ENTAIL label across all the languages. The difference between the accuracy for the CONTRADICT label of French versus Afrikaans and Hindi can entirely be attributed to this sort of confusion. Furthermore, ENTAIL gets quite confused with CONTRADICT.

In Figure 10.4, we also see the greatest language inconsistency with ENTAIL label going towards CONTRADICT across all the languages, though this inconsistency is least in Afrikaans. The inconsistency for CONTRADICT label being predicted as ENTAIL is increasing across resource size of languages from French having the least to Hindi having the highest. Otherwise, the inconsistency across languages is rather low, showing that the XLM-R_{LARGE} model is quite consistent across languages.

In Table 10.8, we can observe that our model on average performs worst for all ENTAIL belonging to Movie category, NEUTRAL and CONTRADICT belonging to City category. In general, our model performs the worst for all hypothesis belonging to the City category possibly because of the involvement of larger table sizes on average and highly numeric and specific hypothesis statements as compared to the rest of the categories. Our models

perform extremely well on all ENTAIL in FoodDrink category because of their smaller table size on average and hypothesis requiring no external knowledge to confirm as compared to CONTRADICT. For ENTAIL our model performs remarkably well on Organization category for French, getting all the hypothesis labels correct. While for NEUTRAL, it performs well for Paintings in French language. Lastly, it performs marginally well for CONTRADICT on Hindi for Organization as compared to the highest performing category for CONTRADICT in English i.e. the Movie category. All language averages perform in the order of their language resource which is expected from Table 10.5.

Table 10.9 depicts a subset of the validation set which has been labeled based on different reasoning mechanisms that the model must employ to categorize the hypothesis correctly. We found the reasoning accuracy scores for 4 languages along with human evaluation score for comparison. Upon observation, we can see that regardless of language, human scores are better than the model we utilize. The variation in language is mostly minimal, but on average our model performs best for English. We notice that for some reasoning types, like Negation and Simple Look-up, humans and the model get no hypothesis right, showing the toughness of the problem. For Numerical based reasoning as well as Coref type reasoning, our model comes very close to human score evaluation. However, overall we are still far from human level performance at TNLI and much scope remains to betterment of models on this task.

10.8 Discussion and Analysis

10.8.1 Extraction versus Translation

One straightforward idea for constructing the multilingual tabular NLI dataset is to extract multilingual tables from Wikipedia in the considered languages. However, this strategy fails in practice for several reasons. For starters, not all articles are multilingual. For example, only 750 of the 2540 tables were from articles available in Hindi. The existence of the same title articles across several languages does not indicate that the tables are identical. Only 500 of the 750 tables with articles in Hindi had infoboxes, and most of these tables were considerably different from the English tables. The tables had different numbers of keys and different value information.

10.8.2 Human Verification versus Human Translation

We selected machine translation with human verification over hiring expert translators for several reasons: (a) Hiring bilingual, skilled translators in multiple languages is expensive and challenging, (b) Human verification is a more straightforward classification task based on semantic similarity; it is also less erroneous compared to translation, (c) By selecting an appropriate verification sample size, we may further minimize the time and effort required for human inspection, (d) A competent translation system has no effect on the classification labels used in inference. As a result, the loss of the semantic connection between the table and the hypothesis is not a significant issue [117], and (e) Minor translation errors have no effect on the downstream NLI task label as long as the semantic meaning of the translation is retained [11, 39, 42, 117].

10.8.3 Usage and Future Direction

The dataset can be used to test benchmarks, multilingual models, and methods for tabular NLI. In addition to language invariance, robustness, and multilingual fact verification, it may well be utilized for reasoning tasks like multilingual question answering [47]. The baselines can also be beneficial to understand models' cross-lingual transferability.

Our current table structure does not generate natural language sentences and hence does not optimize the capabilities of a machine translation model. The representation of tables can be enhanced further by adding Better Paragraph Representation (BPR) from [182]. Additionally, NER handling may be enhanced by inserting a predetermined template name into the sentence post-translation, i.e. extracting a named entity from the original sentence, replacing it with a fixed template entity, and then replacing the named entity with the template post-translation. Multiple experiments, however, would be necessary to identify suitable template entities for replacement, and hence this is left as future work. Another approach is the extraction of keys and values from multilingual Wikipedia pages is also a challenging task and left as future work. Finally, human intervention can enhance the translation quality by either direct human translation or fine-grained post-translation verification and correction.

10.9 Conclusion

We built the first multilingual tabular NLI dataset, namely XINFOTABS, by expanding the INFOTABS dataset with ten different languages. This is accomplished by our novel machine translation approach for tables, which yields remarkable results in practice. We thoroughly evaluated our translation quality to demonstrate that the dataset meets the acceptable standard. We further examined the performance of multiple multilingual models on three validation sets of varying difficulty, with methods ranging from the basic translation-based technique to more complicated language-specific and intermediate task finetuning. Our results demonstrate that, despite the models' success, this dataset remains a difficult challenge for multilingual inference. Lastly, we gave a thorough error analysis of the models to comprehend their cross-linguistic transferability, robustness to language change, and coherence with reasoning.

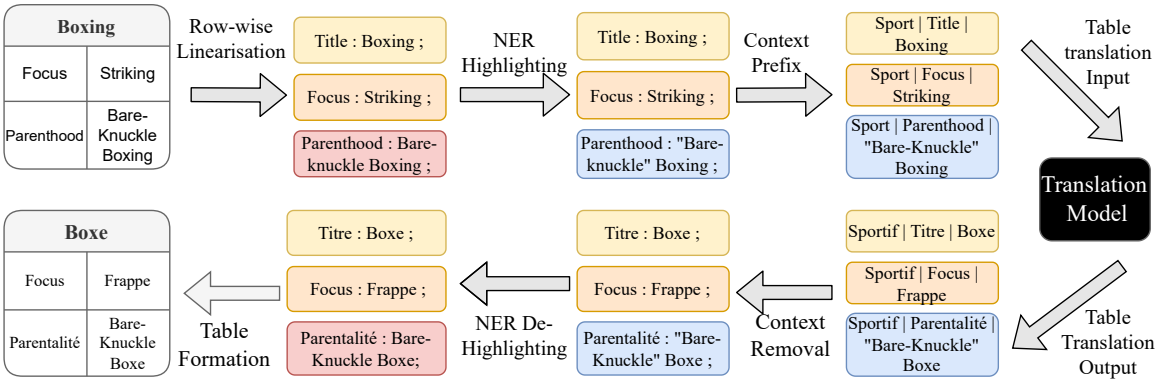
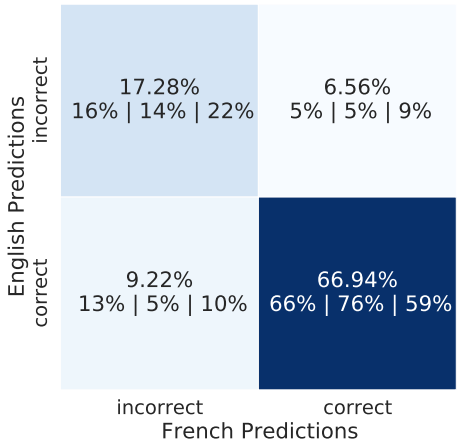
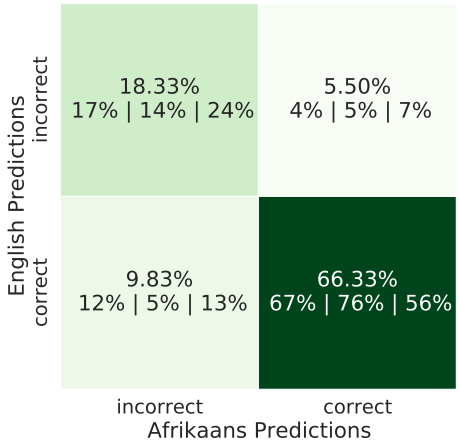


Figure 10.1: Table translation pipeline (§10.4) with premise table "Boxing" (from INFOTABS) translated into French.





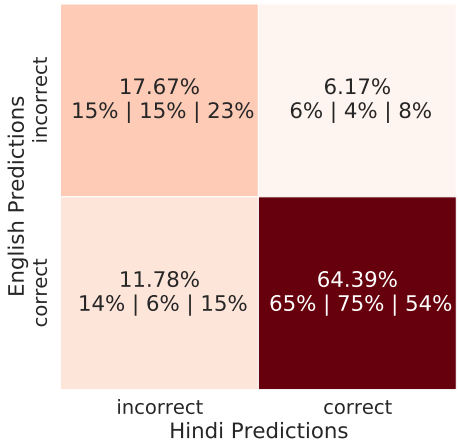


Figure 10.2: Predictions of XLM-RoBERTa for English versus (a) French, (b) Afrikaans, (c) Hindi. The percentage on top in each block represents the average across all three labels with each label percentage given below it in the order of ENTAIL, NEUTRAL and CONTRADICT (cf. §10.7.6).

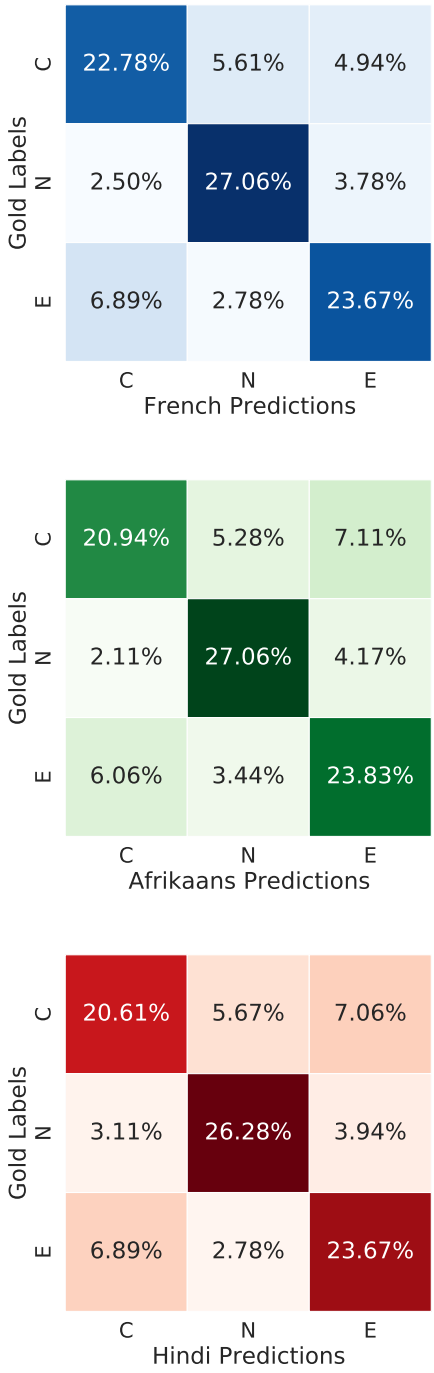


Figure 10.3: Confusion Matrix: Gold Labels versus predictions of XLM-R for (a) French, (b) Afrikaans, (c) Hindi

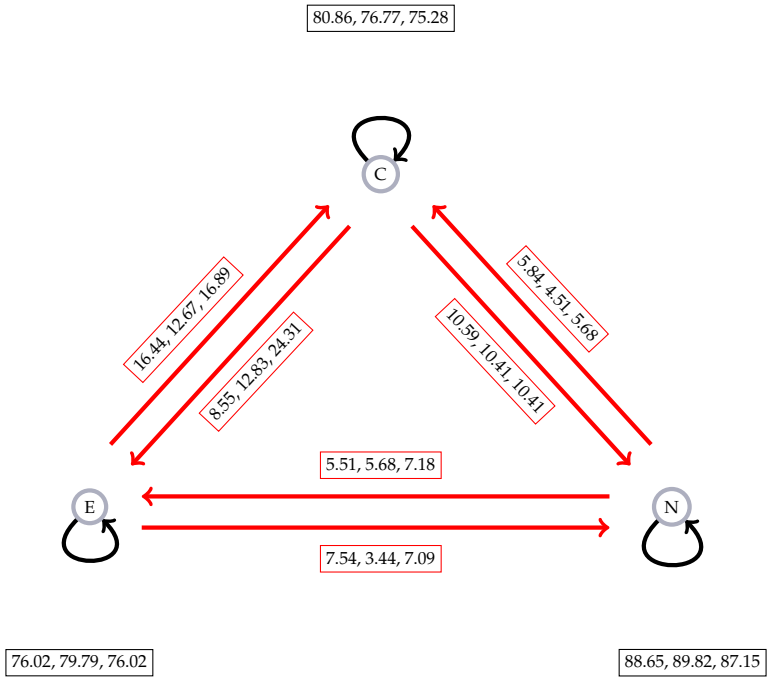


Figure 10.4: Consistency graph for XLM-R (large) predictions of English versus (a) French (b) Afrikaans (c) Hindi in that order respectively.

Table 10.1: Details regarding languages provided in the INFOTABS, from English to Arabic in order of open-source translation resources, refer to OPUS.

Code	Language	Language Family	Script Type	# of Speakers
en	English	Germanic	Latin	1.452 Billion
de	German	Germanic	Latin	134.6 Million
fr	French	Romance	Latin	274.1 Million
es	Spanish	Romance	Latin	548.3 Million
af	Afrikaans	Germanic	Latin	17.5 Million
ru	Russian	Balto-Slavik	Cryllic	258.2 Million
zh	Chinese	Sinitic	Hanzi	1.118 Billion
ko	Korean	Koreanic	Hangul	81.7 Million
hi	Hindi	Indo-Aryan	North-Indic	602.2 Million
ar	Arabic	Semitic	Arabic	274.0 Million

Table 10.2: An example of the XInfoTabS dataset containing English (top-left) and French (top-right) tables in parallel with the hypothesis associated with the table in five languages (below).

Boxing (en)	
Focus	Punching, striking
Olympic sport	688 BC (Ancient Greece), 1904 (modern)
Parenthood	Bare-knuckle boxing
Country of origin	Prehistoric
Also known as	Western Boxing, Pugilism See note.

Boxe (fr)	
Focus	Punching, frappe
Sport olympique	688 av. J.-C. (Grèce ancienne), 1904 (moderne)
Parentalité	Bare-knuckle boxe
Pays d'origine	Préhistorique
Aussi connu sous le nom	Western Boxing, Pugilism Voir note.

Language	Hypothesis	Label
English	The modern form of boxing started in the late 1900's.	CONTRADICT
German	Boxen hat seinen Ursprung als olympischer Sport, der vor Jahrtausenden begann.	CONTRADICT
French	La boxe occidentale implique des punches et des frappes	ENTAIL
Spanish	El boxeo ha sido un evento olímpico moderno durante más de 100 años.	ENTAIL
Afrikaans	Bare-knuckle boks is 'n prehistoriese vorm van boks.	NEUTRAL

Table 10.3: Accuracy scores of the *Table as Struct* strategy on XINFOTABS subsets with $\text{RoBERTa}_{\text{LARGE}}$ model, hypothesis only baseline and majority human agreement results. The first three rows are reproduced from [84].

Model	dev	α_1	α_2	α_3
Human	**79.78**	**84.04**	**83.88**	**79.33**
Hypo Only	60.51	60.48	48.26	48.89
$\text{RoBERTa}_{\text{LARGE}}$	77.61	75.06	69.02	64.61

Table 10.4: Table translation experiment results with Paraphrase Score (PS), Multilingual Paraphrase Score (mPS), BERTScore (BS), Human Evaluation Score (HES), Language Average (LnAvg) and Model Average (MdlAvg). We use the "X | Y" format, where X and Y represent the Table and hypothesis translation score respectively. **Purple** and **Orange** signifies the language average score of the model selected for table and hypothesis translation respectively.

Model	Metric	de	fr	es	af	ru	zh	ko	hi	ar	MdlAvg
MarianMT	PS	95 | 96	93 | 95	93 | 96	83 | 88	81 | 87	75 | 85	N.A.	56 | 55	60 | 79	80 | 85
	mPS	92 | 95	87 | 96	90 | 96	83 | 84	78 | 84	79 | 83	N.A.	65 | 64	66 | 74	80 | 85
	BS	93 | 94	91 | 94	92 | 94	84 | 89	81 | 87	73 | 85	N.A.	63 | 68	64 | 83	80 | 87
	HES	95 | 87	92 | 86	92 | 94	70 | 56	84 | 54	75 | 59	N.A.	40 | 23	58 | 56	76 | 64
	LnAvg	**94** | **93**	**91** | **93**	**92** | **95**	80 | 79	81 | 78	76 | 78	N.A.	56 | 53	62 | 73	79 | 80
mBART50	PS	94 | 96	93 | 95	86 | 87	88 | 92	89 | 87	81 | 87	83 | 82	85 | 82	70 | 77	85 | 87
	mPS	92 | 96	90 | 96	72 | 92	85 | 91	81 | 88	79 | 84	86 | 83	79 | 81	80 | 80	83 | 88
	BS	91 | 94	91 | 93	71 | 88	88 | 93	85 | 89	77 | 86	79 | 85	82 | 86	76 | 83	82 | 89
	HES	93 | 84	91 | 81	82 | 80	89 | 69	87 | 69	76 | 61	76 | 54	79 | 70	71 | 53	83 | 69
	LnAvg	93 | 93	91 | 91	78 | 87	**88** | 86	86 | 83	78 | 80	**81** | 76	**81** | 80	**74** | 73	83 | 83
M2M100	PS	89 | 96	92 | 94	88 | 95	91 | 94	89 | 90	83 | 82	83 | 92	83 | 88	72 | 77	86 | 90
	mPS	88 | 96	88 | 96	88 | 96	84 | 92	83 | 88	80 | 86	84 | 90	81 | 87	78 | 92	84 | 91
	BS	87 | 94	89 | 93	86 | 93	89 | 94	87 | 90	81 | 88	80 | 90	81 | 89	73 | 88	84 | 91
	HES	88 | 85	86 | 86	84 | 86	86 | 83	87 | 74	79 | 72	70 | 82	75 | 73	60 | 51	79 | 77
	LnAvg	88 | 93	89 | 92	87 | 93	88 | **91**	**87** | **86**	**81** | **82**	79 | **89**	80 | **84**	71 | **77**	83 | 87
GoogleTr	PS	91 | 94	94 | 93	92 | 93	96 | 95	79 | 86	80 | 83	87 | 89	90 | 85	60 | 81	85 | 89
	mPS	89 | 94	88 | 94	88 | 94	82 | 87	82 | 86	80 | 86	83 | 87	77 | 80	71 | 81	82 | 88
	BS	87 | 91	89 | 90	88 | 91	88 | 93	77 | 85	78 | 82	82 | 85	87 | 85	63 | 82	82 | 87
	HES	91 | 79	93 | 81	89 | 83	96 | 81	84 | 66	79 | 56	79 | 70	92 | 74	65 | 70	85 | 73
	LnAvg	90 | 90	91 | 90	89 | 90	91 | 89	81 | 81	79 | 77	83 | 83	87 | 81	65 | 79	84 | 84

Table 10.5: Accuracy for baseline tasks on the α_1 set. **Purple** signifies the best task average accuracy, **Orange** signifies the best language average accuracy, **Blue** signifies the best model accuracy. XLM-R_{LARGE} represent XLM-RoBERTa$_{LARGE}$ (XNLI) model.

Train/Test Strategy	Model	en	de	fr	es	af	ru	zh	ko	hi	ar	ModAvg
English Translated Test (§10.7.2)	mBERT$_{BASE}$	-	66	64	65	66	63	63	64	64	59	64
	XLM-R$_{LARGE}$	-	73	73	72	72	72	71	69	70	62	70
	Lang. Avg.	-	**70**	69	69	69	67	67	67	67	61	68
Language Specific Training (§10.7.3)	mBERT$_{BASE}$	67	65	65	63	62	64	63	61	63	57	63
	XLM-R$_{LARGE}$	76	75	74	74	72	71	73	71	71	68	72
	Lang. Avg.	**72**	70	69	68	67	67	68	66	67	63	68
Multiple Language Finetuning Using Only English (§10.7.4A)	mBERT$_{BASE}$	-	64	66	64	64	64	65	63	62	62	64
	XLM-R$_{LARGE}$	-	75	74	75	74	74	73	73	72	69	73
	Lang. Avg.	-	69	**70**	69	69	69	69	68	67	66	**69**
Multiple Language Finetuning Unified Model (§10.7.4B)	mBERT$_{BASE}$	65	64	64	64	64	63	64	62	62	59	63
	XLM-R$_{LARGE}$	76	75	74	75	73	74	74	73	72	70	**74**
	Lang. Avg.	**71**	69	69	70	69	68	69	67	67	65	**69**
English Premise Multilingual Hypothesis (§10.7.5)	mBERT$_{BASE}$	-	63	63	64	62	61	61	59	61	60	61
	XLM-R$_{LARGE}$	-	73	73	73	72	72	73	72	71	68	72
	Lang. Avg.	-	**68**	**68**	**68**	67	67	67	66	66	64	67

Table 10.6: Accuracy for baseline tasks on the α_2 set. **Purple** signifies the best task average accuracy, **Orange** signifies the best language average accuracy, **Blue** signifies the best model accuracy. XLM-R_{LARGE} represent XLM-RoBERTa$_{LARGE}$ (XNLI) model.

Train/Test Strategy	Model	en	de	fr	es	af	ru	zh	ko	hi	ar	Model. Avg
English Translated Test (§10.7.2)	mBERT$_{BASE}$	-	54	53	52	54	52	52	53	52	50	53
	XLM-R_{LARGE}	-	67	66	64	65	65	63	63	63	58	64
	Lang. Avg.	-	60	60	58	60	59	58	58	58	54	59
Language Specific Training (§10.7.3)	mBERT$_{BASE}$	54	54	52	53	50	52	52	51	50	48	52
	XLM-R_{LARGE}	68	66	64	66	63	64	64	64	62	57	64
	Lang. Avg.	61	60	58	60	57	58	58	58	56	53	58
Multiple Language Finetuning Using Only English (§10.7.4A)	mBERT$_{BASE}$	-	53	54	51	53	53	53	52	51	50	52
	XLM-R_{LARGE}	-	66	67	66	66	65	65	65	64	61	65
	Lang. Avg.	-	59	60	58	59	59	59	59	58	55	59
Multiple Language Finetuning Unified Model (§10.7.4B)	mBERT$_{BASE}$	53	51	53	53	52	51	53	50	50	49	52
	XLM-R_{LARGE}	66	64	64	63	64	64	64	63	63	60	64
	Lang. Avg.	60	58	59	58	58	58	58	56	57	54	58
English Premise Multilingual Hypothesis (§10.7.5)	mBERT$_{BASE}$	-	49	53	53	51	49	49	50	47	50	50
	XLM-R_{LARGE}	-	63	65	65	64	65	65	63	63	61	64
	Lang. Avg.	-	56	59	59	57	57	57	57	55	55	57

Table 10.7: Accuracy for baseline tasks on the α_3 set. **Purple** signifies the best task average accuracy, **Orange** signifies the best language average accuracy, **Blue** signifies the best model accuracy. XLM-R_{LARGE} represent XLM-RoBERTa$_{LARGE}$ (XNLI) model.

Train/Test Strategy	Model	en	de	fr	es	af	ru	zh	ko	hi	ar	Model. Avg.
English Translated Test (§10.7.2)	mBERT$_{BASE}$	-	52	53	52	53	53	52	52	52	50	52
	XLM-R_{LARGE}	-	65	65	64	63	64	62	62	61	57	63
	Lang avg	-	58	59	58	58	59	57	57	57	53	58
Language Specific Training (§10.7.3)	mBERT$_{BASE}$	52	50	52	53	50	50	51	48	49	49	50
	XLM-R_{LARGE}	67	65	62	64	62	62	63	60	62	57	62
	Lang avg	60	58	57	58	56	56	57	54	56	53	56
Multiple Language Finetuning Using Only English (§10.7.4A)	mBERT$_{BASE}$	-	52	50	52	52	51	51	49	49	48	50
	XLM-R_{LARGE}	-	65	64	65	62	64	60	63	62	63	63
	Lang avg	-	59	57	58	57	57	56	56	56	54	57
Multiple Language Finetuning Unified Model (§10.7.4B)	mBERT$_{BASE}$	53	50	51	53	50	50	51	47	50	49	50
	XLM-R_{LARGE}	66	64	64	64	63	64	63	62	63	60	63
	Lang avg	60	57	57	58	56	57	57	55	56	54	57
English Premise Multilingual Hypothesis (§10.7.5)	mBERT$_{BASE}$	-	51	50	51	50	50	47	45	48	48	49
	XLM-R_{LARGE}	-	63	63	64	62	62	62	60	61	60	62
	Lang avg	-	57	57	57	56	56	55	54	55	54	56

Table 10.8: Category wise accuracy scores of XLM-R (large) for four languages: namely English (En), French (Fr), Afrikaans (Af) and Hindi (Hi). **Orange** denotes the least score in the column and **Purple** denotes the highest score in the column.

Categories	ENTAIL					NEUTRAL					CONTRADICT				
	En	Fr	Af	Hi	Avg.	En	Fr	Af	Hi	Avg.	En	Fr	Af	Hi	Avg.
Person	79	71	75	73	74	82	81	78	81	81	**59**	67	54	56	59
Musician	**88**	77	78	76	80	87	87	91	82	87	70	69	60	69	67
Movie	**70**	**63**	**57**	63	**63**	85	93	85	87	88	**81**	**76**	**78**	65	**75**
Album	76	76	81	**62**	74	**95**	90	86	**90**	**90**	76	**76**	67	62	70
City	73	58	60	67	65	71	69	**65**	**63**	**67**	67	**54**	**50**	**52**	**56**
Country	74	**61**	65	63	66	74	70	76	76	74	74	72	76	69	73
Painting	83	79	75	67	76	83	**96**	**92**	83	89	71	71	71	71	71
Animal	79	75	79	79	78	75	58	83	67	71	71	75	67	58	68
Food&Drink	**88**	83	75	**88**	**83**	83	79	71	79	78	67	63	58	54	60
Organization	83	**100**	**83**	50	79	**67**	**67**	67	67	**67**	67	67	67	**83**	71
Other	75	73	67	73	72	73	84	84	75	79	76	68	71	62	69
Avg.	79	74	72	69	74	80	79	80	77	79	71	69	65	64	67

Table 10.9: Reasoning wise number of correct predictions of XLM-R (large) for four languages: namely English (En), French (Fr), Afrikaans (Af) and Hindi (Hi) along with human scores for the english dataset.

Reasoning	ENTAIL					NEUTRAL					CONTRADICT				
	H.En	En	Fr	Af	Ko	H.En	En	Fr	Af	Ko	H.En	En	Fr	Af	Ko
Coref	8	6	6	6	4	22	19	19	20	19	13	10	9	7	8
Entity Type	6	5	5	5	5	8	6	6	6	6	6	6	6	4	5
KCS	31	21	19	17	22	21	20	17	19	18	24	18	17	17	20
Lexical Reasoning	5	4	4	4	3	3	2	2	2	1	4	1	1	1	1
Multirow	20	14	11	11	11	16	13	12	13	11	17	15	14	10	13
Named Entity	2	0	0	0	1	2	1	1	1	2	1	1	1	1	1
Negation	0	0	0	0	0	0	0	0	0	0	6	5	5	4	5
Numerical	11	10	7	8	8	3	3	2	3	2	7	5	6	4	4
Quantification	4	2	2	2	2	13	10	10	12	10	6	2	1	2	3
Simple Lookup	3	2	1	2	2	0	0	0	0	0	1	0	1	0	0
Subjective/OOT	6	3	4	4	3	41	37	35	36	37	6	3	4	2	3
Temporal	19	16	12	13	14	11	6	6	6	5	25	18	20	15	19

CHAPTER 11

CONCLUSIONS

In this chapter, we summarize contributions of this dissertation and discuss potential directions for future work.

11.1 Summary

Overall, we address the challenges associated with semi-structured tabular data by proposing effective methods for incorporating knowledge into reasoning models. Two datasets, INFOTABS and AUTO-TNLI, were introduced and used to improve reasoning in the semi-structured, multi-domain, and heterogeneous nature of the premises. The proposed approaches, including simple pre-processing strategies and leveraging structured data knowledge graphs with a novel transformer knowledge Bi-LSTM (TRANSKBLSTM) network, were effective in enhancing model performance and robustness on adversarial tests. Furthermore, the we proposes a trustworthy tabular inference approach to improve model reasoning and interpretability, along with a cost-effective pipeline for translating tables (XINFOTABS) to enable tabular reasoning models to work in multiple languages. This dissertation contributes to advancing the fields of natural language processing by providing effective solutions for reasoning with semi-structured tabular data.

11.2 Looking Forward

In this section, we look at how this work can be further expanded upon.

1. **Challenging Generation Tasks:** To improve table generation tasks, we suggest developing zero-shot models that are tailored to specific table domains. Additionally, we should explore harder tasks, like question generation and true/false inference generation, which require multi-turn interactions. Rather than focusing solely on extractive tasks with explicit reasoning, we should shift our focus towards more abstractive tasks that involve implicit reasoning.

2. **Text to Table:** Generating tables using language models, particularly LLMs, is a promising approach, but there are challenges such as selecting important information and structuring the table correctly [282]. This involve generating a schema from the paragraph category or domain, using LLMs for key phrase extraction and creating a LLM-based zero-shot QA model for the task. Additionally, we recommend using a multimodal LLM that can process text and vision together for truthful generation. We recommend to investigating the use of generative models for more abstract tasks that require implicit reasoning.
3. **Zero-shot Benchmarks:** To improve the performance of existing models and better evaluate their capabilities, new tailored benchmarks are necessary for text-to-table, table question answer, and inference generation. Innovative methods for querying large language models, such as dialogue-based approaches [30, 145, 180], can capture complex information in an organized way and generate more accurate responses. Advanced models that can handle multi-turn tasks, like clarification-based dialogue question answering, are needed, more complex data types [31], requiring complex reasoning and decision-making capabilities and the ability to handle implicit reasoning tasks.
4. **Knowledge Representation and Verification:** To effectively communicate knowledge, tables and text have their respective advantages and limitations. Tables are concise but may contain implicit information, making them challenging to parse. Text is more flexible but can be ambiguous. To address these challenges, we recommend exploring methods to convert between the two forms and integrating both modalities to create a unified representation. Recent research on chart-to-table pretraining has yielded positive results in this area [135, 152, 153]. Additionally, incorporating other sources such as knowledge graphs, captions, and contextual text can provide additional context and enhance the accuracy of the information.
5. **Handling Noisy Information:** Handling noisy ground truth is a difficult challenge that can be addressed through various solutions. One approach is to incorporate multiple sources of information, such as knowledge graphs, captions, and text surrounding the table. Another solution involves learning to generate missing information from pre-trained knowledge while ignoring incorrect data. Type-based con-

straints can be used to filter out problematic existing information, while outdated information may require common sense or knowledge-based multi-expert models for consistency checking. Multi-view expert models, where multiple models are trained on different modalities, can act as fact verifiers to handle ambiguous cases. Finally, multi-turn approaches with subsequent information correction may be more effective than single-turn approaches.

6. **Multiple Information Views:** Future research should explore the integration of multiple modalities such as knowledge bases, databases, images, and videos with text and tables [252]. This can be achieved by incorporating both general and commonsense knowledge graphs to enhance reasoning capabilities. New approaches for querying LLM models, such as dialogue-based methods, should also be considered. Multimodal LLMs that can reason across multiple modalities, convert between them, or utilize pre-training in specific cases are also promising areas for future research.
7. **Domain-specific Information Extraction (IE) from semi-structured tables** has significant potential in unlocking valuable information in domains such as resumes, financial and medical records, clinical reports, and scientific papers. Semantic information in these domains can depend on table headers, making extraction challenging. Tables offer a more concise data representation than free text, which can be unstructured. Key-phrase extraction (require NER, POS tagging, Dependency Parsing, Mutual Bootstrapping for Patterns, and other IE techniques etc.) and is a promising method for semi-structured IE, extracting relevant information from free-form text, such as resumes. Effective methods for domain-specific IE from semi-structured tables can unlock valuable insights, improving decision-making across applications.

There are significant challenges and opportunities in semi-structured data research that require the use of multiple modalities, advanced reasoning, and learning from noisy ground truth. These technologies have the potential to make a significant impact in various domains such as e-commerce, healthcare, and education. Therefore, this area of research is crucial for advancing natural language processing.

11.3 Open Problems

This dissertation opens up new directions for research. Here I list some research questions that can be investigated by expanding upon the findings presented in this work.

(a.) **Dynamic Temporal Reasoning.** Numerous data pieces about an entity evolve and change throughout time. For instance, a city's population, geographical coverage or its official representatives change frequently. ***How do models reason about dynamic, particularly temporally varying information?*** To enable consistent reasoning across time, robust models must consider these temporal variations. I aim to address this challenge by developing methods that leverage time-sensitive language models. Evaluating language model for static temporal reasoning over paragraph and knowledge graph is studied in the past [22, 53, 109, 181, 186, 230, 278, 309].

(b.) **Reducing Information Gaps.** Tables across different languages often have significant information gaps, such as the variation in an entity infoboxes between English and French. ***Ho can models close the information gap across multilingual tables?*** To address this challenge, we can utilize information editing techniques, including information alignment and updating, which can be achieved through the use of large language models. Recently related problems of information editing are explored for article updating [102], news editing [240], headline updation [191], and sentence updation [57, 235].

(c.) **Navigating Multi-modal Information.** My current work involves studying unimodal tables with simple text. It good to expand the semi-structured research to include multimodal tables with text, symbols, images, and complex nested structures. ***How can model reason on complex multimodal tables?*** We can address this question by working with pre-trained models that can analyze both visual and textual information. The model should also account for visual variations, such as highlights, color changes, and font variations. Recently efforts is been made to for similar work specifically on chart-table QA/generation [135, 152, 153], QA on infographicVQA [166, 250], and image-table-text generation [73, 248].

This dissertation research has shown although humans find reasoning on tabular data easy, NLP models, which are primarily designed for unstructured text, struggle. Even when these models appear to make correct inferences, they often do so for the wrong reasons. To address this, models need to incorporate knowledge and focus on the relevant parts of the tabular evidence. By tackling the broader problems of dynamic, multilingual, and multi-modal information in semi-structured data, we can understand reasoning about changing information, and more complex complex data types.

APPENDIX A

QUALITATIVE EXAMPLES

In this section, we provide examples where model is able to predict well after the proposed modifications. We also provide some examples, where model struggles to make the correct prediction after distracting row removal (DRR) modification.

A.1 BPR

Original Premise The Birth name of Eva Mendes are Eva de la Caridad Méndez. Eva Mendes was Born on March 5, 1974 (1974-03-05) (age 44) Miami, Florida, U.S.. The Occupation of Eva Mendes are Actress, model, businesswoman. The Years active of Eva Mendes are 1998 - present. The Partner(s) of Eva Mendes are Ryan Gosling (2011 - present). The Children of Eva Mendes are 2.
Better Paragraph Premise Eva Mendes is a person. The birth name of Eva Mendes is Eva de la Caridad Méndez. Eva Mendes was born on March 5, 1974 (1974-03-05) (age 44) Miami, Florida, U.S.. The occupation of Eva Mendes is Actress, model, businesswoman. The years active of Eva Mendes was on 1998 - present. The partner(s) of Eva Mendes is Ryan Gosling (2011 - present). The number of children of Eva Mendes are 2.
Hypothesis Eva Mendes has two children.

Result and explanation: In this example from α_2, the model predicts Neutral for this hypothesis with Original premise. However, forming better sentences by adding the *"number of children are 2"* (highlighted as green) in case of CARDINAL type for the category PERSON helps the model understand the relation and reasoning behind the children and the number two and arrive at the correct prediction of entailment.

A.2 KG Implicit

Original Premise Janet Leigh is a person. Janet Leigh was born as Jeanette Helen Morrison (1927-07-06) July 6, 1927 Merced, California, U.S. Janet Leigh died on October 3, 2004 (2004-10-03) (aged 77) Los Angeles, California, U.S.. The resting place of Janet Leigh is Westwood Village Memorial Park Cemetery. The alma mater of Janet Leigh is University of the Pacific. The occupation of Janet Leigh are Actress, singer, dancer, author. The years active of Janet Leigh was on 1947-2004. The political party of Janet Leigh is Democratic. The spouse(s) of Janet Leigh are John Carlisle (m. 1942; annulled 1942), Stanley Reames (m. 1945; div. 1949), Tony Curtis (m. 1951; div. 1962), Robert Brandt (m. 1962). The children of Janet Leigh are Kelly Curtis, Jamie Lee Curtis.
Hypothesis A Janet Leigh's career spanned **over** 55 years long.

Hypothesis B Janet Leigh's career spanned **under** 55 years long.

Result and explanation: In this example from α_2, the model without implicit knowledge and the model with implicit knowledge addition predict the correct label on the Hypothesis A. However for Hypothesis B which is an example from α_2, and originally generated by replacing the word "**over**" to word "**under**" in the Hypothesis A and flipping gold label from entail to contradiction, the ealier model which is using artifacts over lexical patterns arrive to predict the original wrong label entail instead of contradiction. On adding implicit knowledge while training, the model is now able to reason rather than relying on artifacts and correctly predicts contradiction. Note, that both hypothesis A and hypothesis B require exactly same reasoning for inference i.e. they are equally hard.

A.3 DRR

Original Premise The pronunciation of Fluorine are (FLOOR-een, -in, -yn) and (FLOR-een, -in, -yn). The allotropes of Fluorine is alpha, beta. The appearance of Fluorine is gas: very pale yellow , liquid: bright yellow , solid: alpha is opaque, beta is transparent. The standard atomic weight are, std(f) of Fluorine is 18.998403163(6). The atomic number (z) of Fluorine is 9. The group of Fluorine is group 17 (halogens). The period of Fluorine is period 2. The block of Fluorine is p-block. The element category of Fluorine is Reactive nonmetal. The electron configuration of Fluorine is [He] 2s 2 2p 5. The electrons per shell of Fluorine is 2, 7. The phase at stp of Fluorine is gas. The melting point of Fluorine is (F-2) 53.48 K (-219.67 °C, -363.41 °F). The boiling point of Fluorine is (F 2) 85.03 K (-188.11 °C, -306.60 °F). The density (at stp) of Fluorine is 1.696 g/L. The when liquid (at b.p.) of Fluorine is 1.505 g/cm 3. The triple point of Fluorine is 53.48 K, 90 kPa. The critical point of Fluorine is 144.41 K, 5.1724 MPa. The heat of vaporization of Fluorine is 6.51 kJ/mol. The molar heat capacity of Fluorine is C p : 31 J/(mol·K) (at 21.1 °C) , C v : 23 J/(mol·K) (at 21.1 °C). The oxidation states of Fluorine is -1 (oxidizes oxygen). The electronegativity of Fluorine is Pauling scale: 3.98. Fluorine was ionization energies on 1st: 1681 kJ/mol, 2nd: 3374 kJ/mol, 3rd: 6147 kJ/mol, (more). The covalent radius of Fluorine is 64 pm. The van der waals radius of Fluorine is 135 pm. The natural occurrence of Fluorine is primordial. The thermal conductivity of Fluorine is 0.02591 W/(m·K). The magnetic ordering of Fluorine is diamagnetic (-1.2×10 -4). The cas number of Fluorine is 7782-41-4. The naming of Fluorine is after the mineral fluorite, itself named after Latin fluo (to flow, in smelting). The discovery of Fluorine is André-Marie Ampère (1810). The first isolation of Fluorine is Henri Moissan (June 26, 1886). The named by of Fluorine is Humphry Davy.

Distracting Row Removal (DRR) The first isolation of Fluorine is Henri Moissan (June 26, 1886). The group of Fluorine is group 17 (halogens). The discovery of Fluorine is André-Marie Ampère (1810). Fluorine was ionization energies on 1st: 1681 kJ/mol, 2nd: 3374 kJ/mol, 3rd: 6147 kJ/mol, (more).

HypothesisFlourine was discovered in the 18th century.

Result and explanation: In this example from the α_3 set, removing distracting rows (sentence except the one in green and blue) definitely helps as there are irrelevant distracting noise and also make premise paragraph long beyond BERT maximum tokenization limits. Before DRR is applied, the model predicts neutral due to a) distracting rows and

b) required information i.e. relevant keys-rows highlighted as green being removed due to maximum tokenization limitation (it's second last sentence). However, after DRR, the prune information retained is only the relevant keys highlighted as green and thus the model is able to predict the correct label.

Negative example: In some examples distracting row removal for DRR remove an relevant rows and hence the model failed to predict correctly on the DRR premise, as shown below:

Original Premise Et in Arcadia ego is a painting. Et in Arcadia ego is also known as Les Bergers d'Arcadie. The artist of Et in Arcadia ego is Nicolas Poussin. The year of Et in Arcadia ego is 1637 - 1638. The medium of Et in Arcadia ego is oil on canvas. The dimensions of Et in Arcadia ego is 87 cm 120 cm (34.25 in 47.24 in). The location of Et in Arcadia ego is Musee du Louvre.
Distracting Row Removal (DRR) Et in Arcadia ego is a painting. The artist of Et in Arcadia ego is Nicolas Poussin. The medium of Et in Arcadia ego is oil on canvas. The dimensions of Et in Arcadia ego is 87 cm 120 cm (34.25 in 47.24 in).
Hypothesis The art piece Et in Arcadia ego is **stored** in the **United Kingdom**

Result and explanation: In this example from the Dev set, the DRR technique used removes the required key *"Location"* (highlighted in red) from the para representation. Hence, the model here predicts neutral as the information regarding where the painting is stored i.e. *"Location"* is removed in the DRR, which the model require for making the correct inference. While in original para, this information is still present and the model is able to arrive at the correct label. Another interesting observation is RoBERTa$_L$ knows Musee du Louvre is a museum in the United Kingdom, showing sign of world-knowledge.

Negative example: In another negative examples distracting row removal for DRR got the relevant rows correct but still the model failed to predict correct label due to spurious correlation, as shown below:

Original Premise Idiocracy is a movie. Idiocracy was directed by Mike Judge. Idiocracy was produced by Mike Judge, Elysa Koplovitz, Michael Nelson. Idiocracy was written by Etan Cohen, Mike Judge. Idiocracy was starring Luke Wilson, Maya Rudolph, Dax Shepard. Idiocracy was music by Theodore Shapiro. The cinematography of Idiocracy was by Tim Suhrstedt. Idiocracy was edited by David Rennie. The production company of Idiocracy is Ternion. Idiocracy was distributed by 20th Century Fox. The release date of Idiocracy is September 1, 2006. The running time of Idiocracy is 84 minutes. The country of Idiocracy is United States. The language of Idiocracy is English. The budget of Idiocracy is $2-4 million. In the box office, Idiocracy made $495,303 (worldwide).

Distracting Row Removal (DRR) Idiocracy was directed by Mike Judge. Idiocracy was produced by Mike Judge, Elysa Koplovitz, Michael Nelson. Idiocracy was written by Etan Cohen, Mike Judge. Idiocracy was edited by David Rennie.
HypothesisIdiocracy was directed and written by the **same** person.

Result and explanation: In this example from the Dev set, the model before DRR predicts the correct label but however on DRR, it predicts incorrect label of neutral. Despite the fact that both the relevant rows require for inference (highlighted in green) is present after DRR. This shows, that the model is looking at more keys than required in the initial case, which are eliminated in the DRR, which force the model to change it prediction. Thus, model is utilising spurious correlation from irrelevant rows to predict the label.

A.4 KG Explicit

Original Premise Julius Caesar was born on 12 or 13 July 100 BC Rome. Julius Caesar died on 15 March 44 BC (aged 55) Rome. The resting place of Julius Caesar is Temple of Caesar, Rome. The spouse(s) of Julius Caesar are Cornelia (84-69 BC; her death), Pompeia (67-61 BC; divorced), Calpurnia (59-44 BC; his death).
Original Premise + KG explicit Julius Caesar died on 15 March 44 BC (aged 55) Rome. **The resting place of Julius Caesar is Temple of Caesar, Rome.** Julius Caesar was born on 12 or 13 July 100 BC Rome. The spouse(s) of Julius Caesar are Cornelia (84-69 BC; her death), Pompeia (67-61 BC; divorced), Calpurnia (59-44 BC; his death). KEY: Died is defined as pass from physical life and lose all bodily attributes and functions necessary to sustain life . KEY: Resting place is defined as a cemetery or graveyard is a place where the remains of dead people are buried or otherwise interred . KEY: Born is defined as british nuclear physicist (born in germany) honored for his contributions to quantum mechanics (1882-1970) . KEY: Spouse is defined as a spouse is a significant other in a marriage, civil union, or common-law marriage .
Hypothesis Julius Caesar was buried in Rome.

Result and explanation: In this example from α_2, the model without explicit knowledge predicts neutral for the hypothesis as it is not able to infer that **resting place** is where people are **buried**, so it predicts neutral as it implicitly lack buried key understanding. On explicit KG addition (highlighted as blue+ green), we add the definition of resting place to be the place where remains of the dead are buried (highlighted as green). Now the model uses this extra information (highlighted as green) plus the original key related to death (highlighted in **bold**) to correctly infer that the statement Caesar is buried in Rome is entailed.

Table A.1: Prediction after BPR. Here, + represents the change with respect to the previous row.

Premise	Label
Human Label (Gold)	Entailed
Original Premise	Neutral
+BPR	Entailed

Table A.2: Prediction on Hypothesis A. Here, + represents the change with respect to the previous row.

Premise	Label
Human Label (Gold)	Entailed
Original Premise	Entailed
+ KG implicit	Entailed

Table A.3: Prediction on Hypothesis B (from α_2). Here, + represents the change with respect to the previous row.

Premise	Label
Human Label (Gold)	Contradiction
Original Premise	Entailed
+ KG implicit	Contradiction

Table A.4: Prediction after DRR. Here, + represents the change with respect to the previous row.

Premise	Label
Human Label (Gold)	Contradiction
Original Premise	Neutral
+DRR	Contradiction

Table A.5: Prediction after DRR. Here, + represents the change with respect to the previous row.

Premise	Label
Human Label (Gold)	Contradiction
Original Premise	Contradiction
+DRR	Neutral

Table A.6: Prediction after DRR. Here, + represents the change with respect to the previous row.

Premise	Label
Human Label (Gold)	Entailed
Original Premise	Entailed
+DRR	Neutral

Table A.7: Prediction after KG explicit addition. Here, + represents the change with respect to the previous row.

Model	Label
Human Label (Gold)	Entailed
Original Premise	Neutral
+ KG explicit	Entailed

APPENDIX B

KNOWLEDGE INFOTABS TRANSKBLSTM

B.1 Hypothesis Attention Module

In Hypothesis attention module, we calculate hypothesis relation values by normalizing R_{ijk} with respect to row-axis(2), to generate $R_{ik}^{hyp} \in \mathbb{R}^{m \times k}$ which is the average hypothesis relation for every premise word.

$$R_{ik}^{hyp} = \sum i = 1^{n} \frac{R_{ijk}}{n}$$

We reduce the dimension by applying the dot product attention.

$$R_{ik}^{r} = F_{H}^{r}(R_{ik}^{hyp}) \in \mathbb{R}^{m \times l_k}$$

F_{N}^{r} can again be a single layer neural network. We then use the Hypothesis attention head to highlight the importance of the hypothesis and its relations to the premise. The context-aware premise hidden state p^s is used as queries, the hypothesis hidden state is used as keys, and reduced hypothesis premise relation values are used. The attention function can be defined as follows:

$$\text{Attention}(p^s, h^s, R_{ik}^{r}) = \text{softmax}(\frac{p^s h^{s\,T}}{\sqrt{l}}) R_{ik}^{r}$$

Then the multi-head attention is as follows:

$$\begin{aligned} p_h^{att} &= \text{MH}(p^s, h^s, R_{ik}^{r}) \\ &= \text{Concat}(\text{head}_1, ..., \text{head}_h) W^o \end{aligned}$$

where, head$i$ = Attention$(p^s W_i^q, h^s W_i^k, R^r ikW_i^v)$ and W_i^q, W_i^k, and W_i^v are projection matrices and i is the number of attention heads. The output $p_h^{att} \in \mathbb{R}^{m \times l_k}$ is an attention-weighted context matrix measuring the importance of premise and relations to each of the hypothesis. We calculate $p_h^{att} \in \mathbb{R}^{m \times l_k}$, attention-weighted context matrix measuring the impor-

tance of premise and relations to each of the hypothesis. We also extract H^{att}, the attention weights of the hypothesis multi-head attention.

B.2 Qualitative Examples

Tables B.1, B.2, B.3, B.4, and B.5 present examples to supplement the results presents in Section 5.4.

B.3 Knowledge Relations to Sentence Conversion

We create templates to convert knowledge relations in ConceptNet & WordNet to natural language sentences. These templates resemble natural English text, which can be encoded using pretrained language models. The templates can be seen in Table B.7.

B.4 Domain Analysis

To understand the models performance across tabular domains (i.e. RQ3(b)), we analyse domain-wise table category results. We evaluate the twelve major categories contained in the INFOTABS datasets. All remaining categories are grouped together in the *"Other"* category. Table summarizes the performance of models (trained with 2% and 5% INFOTABS train data)[1] on the INFOTABS development set across several categories.

As the supervision increases from 1% to 10%, we observe an increasing accurate prediction trend across the categories. Our proposed model shows significant improvements in *"Musician"* and *"Sports"* categories. We attribute these huge gains to two main reasons: (a) . Under minimal supervision, knowledge relations enable the model to concentrate on relevant context, thus helping in establishing premise rows and hypothesis tokens connections. For example refer to Table B.1. (b) and the acquisition of additional knowledge enhances the models' overall world knowledge and common sense reasoning capability. E.g. in the Table 5.1, the understanding of the *California* is located at the *coast*.

Additionally, we observe that our proposed model performs poorly in a few categories. This part comprises instances from *"Album"*, *"Food & Drinks"*, and *"University"*. This can be attributed to the noisy addition of knowledge. Sometimes knowledge relations give out the relational context that might not be needed. For example refer to Table B.2 in

[1]For details results on other percentages refer to Appendix §B.4 Table B.8.

Appendix §B.2. Additional knowledge filtering may be addressed in future studies. For domain analysis results of models trained on 2% and 5% training data, refer to Table B.8.

B.5 Limited Supervision

We present detailed results on limited supervision experiments. All the reported numbers are average over three seed runs with a standard deviation of 0.233 (w/o KG), 0.49 (KG Explicit), 0.5 (Tok-KTrans), and 0.30 (Trans-KBLSTM). All the improvements are statistically significant with $p < 0.05$ of one-tailed Student t-test.

B.6 Additional Results Reasoning Analysis

Table B.9 shows the results of our experiments, where we train under limited supervision setting on 1%, 3%, 5% and 10% data. Table B.10 detailed results of performance across reasoning keys for models trained on 1%, 3%, 5% and 10% data.

B.7 Training and Hyperparameters Details

Trans-KBLSTM is implemented in PyTorch [196] using Huggingface [281] implementation of RoBERTa [158]. We pretrain the transformer components on MultiNLI dataset [280] for fair comparison with the Knowledge-INFOTABS baseline of [182]. We use AdaGrad optimizer [56] with an initial learning rate of 1e-4 for RoBERTa and 1e-3 for non-RoBERTa i.e. LSTM parameters with a scheduler. The batch size is slected from {3, 4, 5}. All the hyper-parameters are fined tuned on the development set of INFOTABS. For more details about hyperparameters refer to the Table B.11.

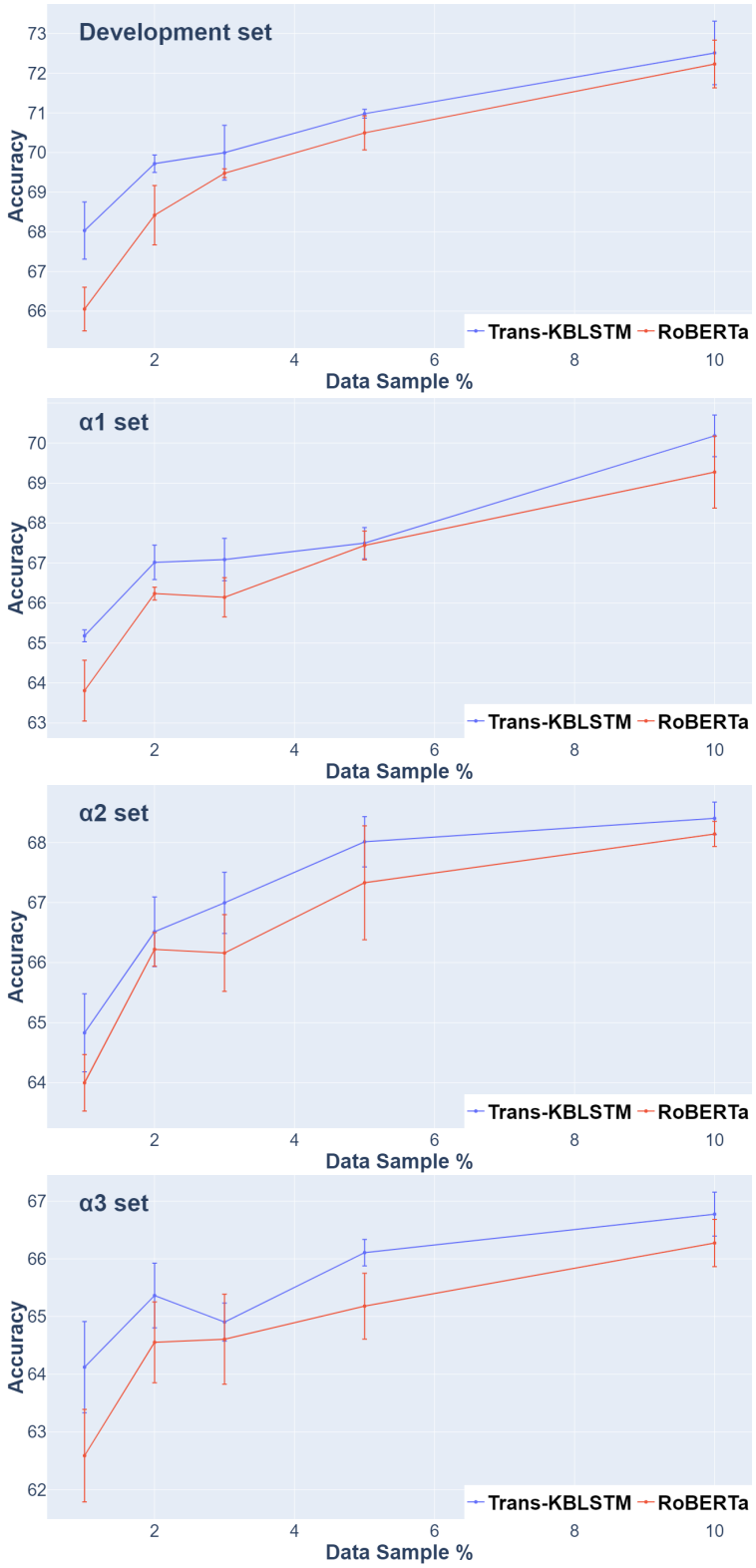


Figure B.1: The figures show error bar plots of limited supervision training on 1, 2, 3, 5, 10 and 15% of data. for Trans-KBLSTM and RoBERTa baseline. We notice that the error overlap increases with increase in supervision. The improvements are higher under low-data regimes.

Table B.1: In the absence of knowledge, the model is unable to understand the word *twenties* and concludes that the information is not present in the text. However, addition of knowledge re-enforces the connection between *age* and *twenties* thereby producing correct label.

Joe Budden Premise	
Premise	Joe Budden was Born on (1980-08-31) August 31, 1980 (age 38) in New York, New York. The Origin of Joe Budden are Jersey City, New Jersey. The Years active of Joe Budden are 1999-present. The Labels of Joe Budden are Mood Muzik, EMPIRE (current), Desert Storm, Def Jam, Amalgam Digital, and E1 (former)
Hypothesis	Joe Budden started his career in his twenties.
Focused Relation	age $\xleftrightarrow{Co-Hyponym}$ twenties
Gold Label	Contradiction
Prediction	
RoBERTa	Neutral
Trans-KBLSTM	Contradiction

Table B.2: The baseline prediction correctly predicts the gold label. Our proposed model gets confused with semantically irrelevant relations and hence concludes the statement as contradiction.

Crooked Teeth Premise	
Premise	The Released of Crooked Teeth are May 19, 2017. The Studio of Crooked Teeth are Steakhouse Studios, North Hollywood, CA. The Genre of Crooked Teeth are Hard rock, nu metal, and rap rock. The Label of Crooked Teeth are Eleven Seven.
Hypothesis	The album Crooked Teeth took over a year to make.
Focused Relation	genre $\xleftrightarrow{Co-Hyponym}$ make —— metal $\xrightarrow{RelatedTo}$ make —— rap $\xrightarrow{Hypernym}$ make
Gold Label	Neutral
Prediction	
RoBERTa	Neutral
Trans-KBLSTM	Contradiction

Table B.3: The inference of the hypothesis requires the model to focus on 1^{st} and 2^{nd} sentences at the same time. The original model gets confused due to mention of *age 69* and *young* and concludes contradiction. The focused relations develop appropriate connections to the first two sentences and enable better understanding to the model.

Jeff Bridges Premise	
Premise	The Born of Jeff Bridges are December 4, 1949 (age 69) Los Angeles, California, U.S.. The Years active of Jeff Bridges are 1951-present. The Children of Jeff Bridges are 3. The Family of Jeff Bridges are Beau Bridges (brother), and Jordan Bridges (nephew).
Hypothesis	Jeff Bridges started his career as a young child.
Focused Relations	born $\xrightarrow{RelatedTo}$ young
	born $\xrightarrow{RelatedTo}$ child
	child $\xrightarrow{RelatedTo}$ age
	active $\xrightarrow{Co-Hyponym}$ child
Gold Label	Entailment
Prediction	
RoBERTa	Contradiction
Trans-KBLSTM	Entailment

Table B.4: The inference of the given hypothesis requires the knowledge of Synonymy between *Corn* and *Maize*.

Chibuku Shake Premise	
Premise	The Type of Chibuku Shake shake are Opaque Beer. The Alcohol by volume of Chibuku Shake shake are 3.3% to 4.5%. The Colour of Chibuku Shake shake are Tan-pink to white. The Ingredients of Chibuku Shake shake are Sorghum, and Maize.
Hypothesis	Corn is an ingredient found in a Chibuku Shake.
Focused Relations	corn $\xleftrightarrow{Synonym}$ maize
Gold Label	Entailment
Prediction	
RoBERTa	Entailment
Trans-KBLSTM	Entailment

Table B.5: The focused external knowledge relation connects the *Monarchy* in premise to *Kingdom* in hypothesis.

Hashemite Kingdom of Jordan Premise	
Premise	The Legislature of Hashemite Kingdom of Jordan are Parliament. The Religion of Hashemite Kingdom of Jordan are 95% Islam (official), 4% Christianity, and 1% Druze, Baha'i. The Government of Hashemite Kingdom of Jordan are Unitary parliamentary constitutional monarchy. The Monarch of Hashemite Kingdom of Jordan is Abdullah II.
Hypothesis	Hashemite Kingdom of Jordan does not have any democracy.
Focused Relation	Kingdom $\overset{IsA}{\longleftrightarrow}$ Monarch
Gold Label	Contradiction
Prediction	
RoBERTa	Neutral
Trans-KBLSTM	Contradiction

Table B.6: Accuracy (%) across different categories observed in the Development set (Others (¡10%) includes the categories, University, Awards, Event, Book and Aircraft), trained on 1%, 3% and 5% samples of the data. **w/o KG** represents RoBERTa and **w KG** represents Trans-KBLSTM model.

Category	1%		3%		10%	
	w/o KG	w KG	w/o KG	w KG	w/o KG	w KG
Album	65.87	65.87	73.81	**76.98**	77.78	73.02
Animal	60.49	**66.67**	75.31	66.67	67.9	**72.84**
City	64.05	**64.71**	56.21	**61.44**	63.4	**64.71**
Country	56.48	54.63	56.48	55.56	60.19	**62.96**
Food & Drinks	69.44	**70.83**	72.22	**73.61**	83.33	79.17
Movie	61.11	**63.89**	63.89	63.89	70	**73.89**
Musician	62.57	**69.88**	73.1	**74.56**	75.73	**76.9**
Organization	61.11	58.33	55.56	**66.67**	69.44	**72.22**
Painting	80.25	80.25	75.31	**77.78**	77.78	**80.25**
Person	57	**62.96**	62.35	**67.28**	74.9	**75.72**
Sports	65.08	**73.02**	61.9	**71.43**	68.25	**69.84**
Others	63.89	**65.28**	66.67	**70.84**	63.89	61.11
TOTAL	62	**65.83**	65.88	**68.61**	72.27	**73.22**

Table B.7: ConceptNet and Wordnet Relations with their Natural language templates.

KB Relation	Natural Language	KB Relation	Natural Language
Antonym	is opposite of	Co-Hyponym	is co-hyponym of
AtLocation	is at location	CapableOf	is capable of
Causes	causes	CausesDesire	causes desire to
CreatedBy	is created by	DefinedAs	is defined as
DerivedFrom	is derived from	Desires	desires
DistinctFrom	is distinct from	Entails	entailes
EtymologicallyDerivedFrom	is etymologically derived from	HasA	has a
ExternalURL	external url	FormOf	is a form of
EtymologicallyRelatedTo	is etymologically related to	HasContext	has context
HasLastSubevent	has last subevent	HasPrerequisite	has prerequisite
HasProperty	has property	HasSubevent	has subevent
InstanceOf	is an instance of	IsA	is a
LocatedNear	is located near	MadeOf	is made of
MannerOf	is manner of	MotivatedByGoal	is motivated by goal
NotCapableOf	is not capable of	NotDesires	does not desire
NotHasProperty	does not have property	PartOf	is part of
ReceivesAction	receives action	RelatedTo	is related to
SimilarTo	is similar to	SymbolOf	is a symbol of
Synonym	is same as	UsedFor	is used for
dbpedia/capital	has capital	dbpedia/field	has field
dbpedia/genre	has genre	dbpedia/genus	has genus
dbpedia/influencedBy	is influenced by	dbpedia/knownFor	is known for
dbpedia/language	has language	dbpedia/leader	has leader
dbpedia/occupation	has occupation	dbpedia/product	has product
Hypernym	is hypernym of	Hyponym	is hyponym of
HasFirstSubevent	has first subevent		

Table B.8: Accuracy (%) across different categories observed in the Development set (Others (¡10%) includes the categories, University, Awards, Event, Book and Aircraft), trained on 2% and 5% samples of the data. **w/o KG** represents RoBERTa baseline and **w KG** represents Trans-KBLSTM.

Category	2%		5%	
	w/o KG	w KG	w/o KG	w KG
Album	68.25	67.46	72.22	**73.81**
Animal	65.43	64.20	72.84	69.14
City	55.56	**58.17**	60.13	**61.44**
Country	58.33	**62.96**	61.11	**68.52**
Food&Drinks	69.44	66.67	75.00	73.61
Movie	58.33	**65.00**	65.56	65.56
Musician	68.42	**71.64**	71.35	**76.32**
Organization	58.33	**61.11**	66.67	66.67
Painting	66.67	59.26	75.31	**76.54**
Person	61.32	60.49	68.72	67.08
Sports	66.67	**69.84**	61.90	**68.25**
Others	62.50	**66.67**	63.89	**65.28**
TOTAL	63.11	**64.44**	68.22	**69.50**

Table B.9: Shows the results of of our experiments, where we train under limited supervision setting. **w/o KG** Original RoBERTa baseline, **KG Explicit** KG-Explicit knowledge addition, **Tok-KTrans** Token appended transformers, **Trans-KBLSTM** Proposed model. We train these models on data samples 1, 2, 3, 5, 10, 15, 20, 25, 30, 50, 100 %s.

Model	% Train	Dev	α1	α2	α3	% Train	Dev	α1	α2	α3
w/o KG	1%	66.05	63.81	64.00	62.59	2%	68.42	66.24	66.22	64.55
KG Explicit		65.15	63.22	62.24	60.63		66.70	65.07	63.77	62.11
Tok-KTrans		63.57	61.96	58.83	59.18		67.74	66.59	62.46	62.78
Trans-KBLSTM		**68.03**	**65.18**	**64.83**	**64.12**		**69.72**	**67.02**	**66.51**	**65.36**
w/o KG	3%	69.48	66.14	66.16	64.61	5%	70.50	67.44	67.33	65.18
KG Explicit		68.12	66.05	64.85	62.85		68.78	66.65	65.20	63.74
Tok-KTrans		67.52	66.57	63.98	64.07		69.44	67.31	65.14	63.53
Trans-KBLSTM		**70.00**	**67.09**	**67.00**	**64.90**		**70.98**	**67.50**	**68.01**	**66.11**
w/o KG	10%	72.23	69.27	68.14	66.27	15%	72.92	70.27	68.46	66.66
KG Explicit		70.68	68.77	67.07	64.70		72.05	70.16	67.37	65.05
Tok-KTrans		71.24	69.79	65.25	65.29		72.47	70.94	66.68	65.20
Trans-KBLSTM		**72.51**	**70.18**	**68.40**	**66.77**		**73.61**	**70.96**	**68.90**	**67.29**
w/o KG	20%	74.09	71.25	69.31	67.68	25%	74.50	72.25	68.90	67.53
KG Explicit		72.70	70.99	67.89	65.55		74.46	72.32	68.61	66.91
Tok-KTrans		73.05	70.77	67.72	65.94		74.44	72.79	68.22	66.83
Trans-KBLSTM		**74.29**	**72.16**	**69.77**	67.29		**75.09**	**73.20**	**69.57**	**68.18**
w/o KG	30%	74.70	72.86	69.61	67.55	50%	75.93	73.79	69.59	67.90
KG Explicit		74.83	72.26	68.69	66.89		75.99	74.05	70.36	68.51
Tok-KTrans		74.17	73.96	68.03	66.63		78.44	76.38	70.66	70.38
Trans-KBLSTM		**75.57**	**74.25**	**69.62**	**67.57**		**76.71**	**74.86**	**70.68**	**68.93**
w/o KG	100%	77.30	76.44	70.49	69.05					
KG Explicit		78.97	77.84	**71.13**	69.58					
Tok-KTrans		78.17	76.19	70.75	69.77					
Trans-KBLSTM		**79.73**	**78.92**	**71.62**	**70.21**					

Table B.10: The above numbers represent accuracy on development dataset across different reasoning types with varying percentage of data. The third number indicates the number of examples corresponding to the reasoning type and label.

Data (%)	Reasoning Keys	Entailment B.L	Entailment KtLSTM	.	Neutral B.L	Neutral KtLSTM	#	Contradiction B.L	Contradiction KtLSTM	#
1%	KCS	64.52	**70.97**	31	85.71	85.71	21	50.00	**62.50**	24
	coref	50.00	**62.50**	8	81.82	68.18	22	30.77	15.38	13
	entitytype	83.33	83.33	6	87.50	87.50	8	50.00	50.00	6
	lexicalreasoning	40.00	**60.00**	5	33.33	33.33	3	25.00	25.00	4
	multirowreasoning	60.00	**75.00**	20	68.75	**75.00**	16	52.94	47.06	17
	nameidentity	0.00	0.00	2	0.00	**100.00**	2	100.00	100.00	1
	negation	0.00	0.00	0	0.00	0.00	0	66.67	**83.33**	6
	numerical	63.64	54.55	11	66.67	**100.00**	3	42.86	42.86	7
	quantification	25.00	25.00	4	100.00	92.31	13	16.67	16.67	6
	subjectiveoot	33.33	33.33	6	75.61	**80.49**	41	50.00	50.00	6
	temporal	73.68	**78.95**	19	45.45	45.45	11	56.00	**60.00**	25
3%	KCS	67.74	**83.87**	31	66.67	**80.95**	21	75.00	70.83	24
	coref	37.50	**50.00**	8	54.55	**63.64**	22	53.85	53.85	13
	entitytype	50.00	50.00	6	62.50	**87.50**	8	66.67	50.00	6
	lexicalreasoning	60.00	**80.00**	5	33.33	**66.67**	3	75.00	75.00	4
	multirowreasoning	60.00	**70.00**	20	56.25	**68.75**	16	76.47	76.47	17
	nameidentity	50.00	**100.00**	2	100.00	100.00	2	100.00	100.00	1
	negation	0.00	0.00	0	0.00	0.00	0	100.00	100.00	6
	numerical	54.55	**81.82**	11	66.67	66.67	3	71.43	71.43	7
	quantification	75.00	75.00	4	69.23	**76.92**	13	66.67	66.67	6
	subjectiveoot	50.00	50.00	6	65.85	**80.49**	41	66.67	66.67	6
	temporal	47.37	**63.16**	19	54.55	**72.73**	11	64.00	40.00	25
5%	KCS	87.10	83.87	31	71.43	**90.48**	21	66.67	62.50	24
	coref	75.00	62.50	8	68.18	**81.82**	22	30.77	30.77	13
	entitytype	83.33	83.33	6	87.50	87.50	8	83.33	83.33	6
	lexicalreasoning	60.00	**80.00**	5	33.33	**66.67**	3	50.00	50.00	4
	multirowreasoning	85.00	85.00	20	68.75	**81.25**	16	58.82	**76.47**	17
	nameidentity	100.00	100.00	2	50.00	**100.00**	2	100.00	0.00	1
	negation	0.00	0.00	0	0.00	0.00	0	100.00	66.67	6
	numerical	72.73	**90.91**	11	100.00	100.00	3	71.43	**85.71**	7
	quantification	75.00	50.00	4	92.31	**100.00**	13	33.33	16.67	6
	subjectiveoot	66.67	33.33	6	73.17	**87.80**	41	50.00	50.00	6
	temporal	94.74	84.21	19	36.36	**63.64**	11	56.00	52.00	25
10%	KCS	74.19	**80.65**	31	95.24	90.48	21	70.83	70.83	24
	coref	50.00	**75.00**	8	77.27	77.27	22	46.15	23.08	13
	entitytype	66.67	**83.33**	6	87.50	87.50	8	100.00	83.33	6
	lexicalreasoning	80.00	80.00	5	66.67	66.67	3	25.00	**75.00**	4
	multirowreasoning	80.00	80.00	20	81.25	81.25	16	76.47	70.59	17
	nameidentity	50.00	50.00	2	100.00	100.00	2	100.00	100.00	1
	negation	0.00	0.00	0	0.00	0.00	0	83.33	**100.00**	6
	numerical	81.82	**100.00**	11	100.00	100.00	3	71.43	71.43	7
	quantification	50.00	50.00	4	84.62	**92.31**	13	33.33	33.33	6
	subjectiveoot	33.33	**50.00**	6	82.93	**87.80**	41	50.00	33.33	6
	temporal	78.95	**89.47**	19	63.64	63.64	11	68.00	68.00	25

Table B.11: Enlists the hyperparameters used while training the baselines and proposed model on INFOTABS.

Hyperparameter	Value
LSTM Max Length	200
LSTM layers	2
LSTM learning rate	1e-3
LSTM Hidden state size	128
Word Embedding Dimension	300
RoBERTa Hidden state size	768
RoBERTa learning rate	1e-4
# Attention heads	4
Embedding Spatial Dropout	0.3
Dropout (Final classification)	0.2

APPENDIX C

SYSTEMATIC PROBE ANNOTATION DETAILS

C.1 Manual Probing Human Annotation

We ask human annotators on Amazon Mechanical Turk to mark relevant rows for each given table and hypothesis. We use the development and test sets (4×1800 pairs) of the INFOTABS dataset for this purpose. The templates with detailed instruction and examples for the annotation is provided at https://tabprobe.github.io/.

Each HIT consists of three table-sentence pairs from the same table. Annotators are asked to mark the rows which are relevant to the given hypothesis. Thus, each row was considered as an independent option to select/not select, making this a multi-label selection annotation problem. Since many hypothesis sentences (especially ones with neutral labels) use out-of-table information, we add an additional option of selecting out-of-table (OOT) information, which is marked only when information not present in the table is used in the hypothesis. We do not provide the labels to annotators to avoid any bias arising due to correlation between the NLI label and row selections. One example of such bias is always selecting/not selecting the OOT depending on the neutral/non-neutral labels.

We repeated the task 5 times for every hypothesis. For every row, we took the most common label (relevant or not) as the ground truth. We also follow standard approaches to improve annotation quality: We employed high-quality master annotators, released examples in batches (50 batches including 2 pilot batches, each with 50 HITs, where each HIT has 1 table with 3 sentences), blocked bad annotators and rewarded bonuses to good ones.We used the degree of agreement between the annotator and the majority to identify good and bad annotators.[1] Furthermore, after every batch, we re-annotated examples with poor consensus, and removed HITs corresponding to 452 pairs (3.6%) and 22 annotators due to poor overall consensus. Our annotation scheme ensures that samples have at least

[1]The supplement at https://tabprobe.github.io/ has more details.

5 diverse annotations each, as shown in Figure C.1. We ensure that annotators get paid above minimum wage by timing the hits ourselves. The final cost of the whole annotation was $\$1,750$ including the pilots, re-annotation, bonuses and any other expenses.

Table C.1 shows inter-annotator agreement via macro average F1-score w.r.t majority for all the annotations, before and after post-annotation data cleaning. Detailed analysis for each kappa bucket is shown in Table 7.2.

C.1.1 Pairwise Annotator Agreement

Table C.3 shows details for inter-annotator agreement for the relevant row annotation task. We obtain an average F1-score of 89.0%(macro) and 88.6%(micro) for all our experiments.

Figure C.2 shows fine-grained agreement i.e., the percentage of examples with varying precision and recall. Figure C.1 shows the percentage of examples with number of annotations (distinct annotators). Figure C.3 shows the percentage of examples pairs(y-axis) verses number of annotations with the exact same labeling for the relevant rows(x-axis).

C.1.2 Human Bias with Information

Our annotation allows us to ask: *Where the annotators who wrote the hypotheses in the original data biased towards specific kinds of rows?* We found that INFOTABS predominantly has hypotheses which focus more on a some rows from the premise table than others; on the other-side, some row keys are completely ignored. Of course, not all keys occur equally in all tables, and therefore, we only consider keys that reoccur substantially and frequently in the INFOTABS premise tables (i.e. not rare/uncommon keys). To account for varying appearance frequency of keys in the tables, in all the analyses we only consider keys which appear $\geq$180 times in the training set.

Figure C.4 shows the human bias problem with the INFOTABS annotations. For ENTAIL and CONTRADICT labels rows keys such as *born*, *died*,*years active*, *label*, *genres*, *release date*, *spouse*, *occupation* are overused. Similarly for the NEUTRAL label rows keys such as *born*, *associated act*, and *spouse* are overused. On the other-side, for ENTAIL, NEUTRAL and CONTRADICT labels rows keys such as *website*, *birth-name*, *type*, *origin*, *budget*, *language*, *edited by*, *directed by*, *running time*, *nationality*, *cinematography* are under used despite frequently appearing in the tables. Some keys such as *years active*, *label*, *genres*, *release date*, *spouse*, and

cinematography are underused only for the NEUTRAL label.

Although INFOTABS discusses the presence of hypothesis bias because of workers knowingly writing similar hypotheses, it does not discuss the possible bias based on usage of the premise rows (keys) and its possible effects on model prediction. We suspect such bias occurs because, during NLI data creation, annotators excessively use keys with *numerical* and *temporal* values to create a hypothesis, as that makes samples easier and faster to create. One possible approach to handle this data bias would be to force NLI data creation annotators to write hypotheses using randomly selected parts of the premises.

C.2 Multi Row Reasoning

We also analyse the proportion of annotated examples using more than one row for a given hypothesis, i.e. multiple row reasoning. As shown in Figure C.5 around 54% and 37% examples have only one or two rows relevant to the hypothesis, respectively. Furthermore, we find that annotators mark OOT mostly for the neutral labels i.e. 71% compared to 5% for combined ENTAIL and CONTRADICT labels.[2]. We also find a very negligible $< 0.25\%$ of examples have zero relevant rows; we suspect this might be because of annotation ambiguity or the hypothesis being a universal factual statement. Overall our annotation analysis verifies the claim made by INFOTABS [84] in their reasoning analysis.

[2]The marking of OOT for ENTAIL and CONTRADICT is mostly attributed to use of commonsense knowledge.

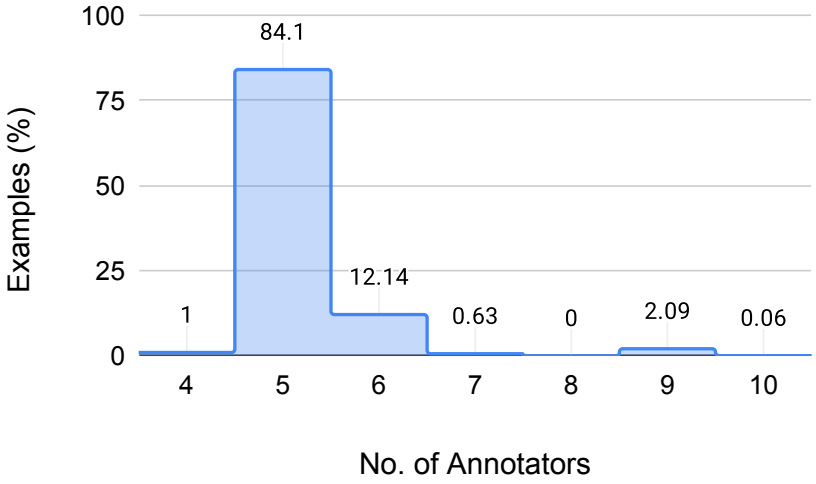


Figure C.1: Percentage of example pairs verses number of diverse annotations for the relevant rows.

		Recall					
		< 50	< 60	< 70	< 80	< 90	< 100
Prec	< 40	0	1	1	1	1	2
	< 50	0	2	2	4	4	5
	< 60	0	3	3	8	8	11
	< 70	0	3	5	11	11	20
	< 80	0	4	7	19	21	40
	< 90	0	5	8	21	25	46
	< 100	1	6	12	32	41	100

Figure C.2: Fine-grained agreement analysis showing the percentage of examples with given precision and recall.

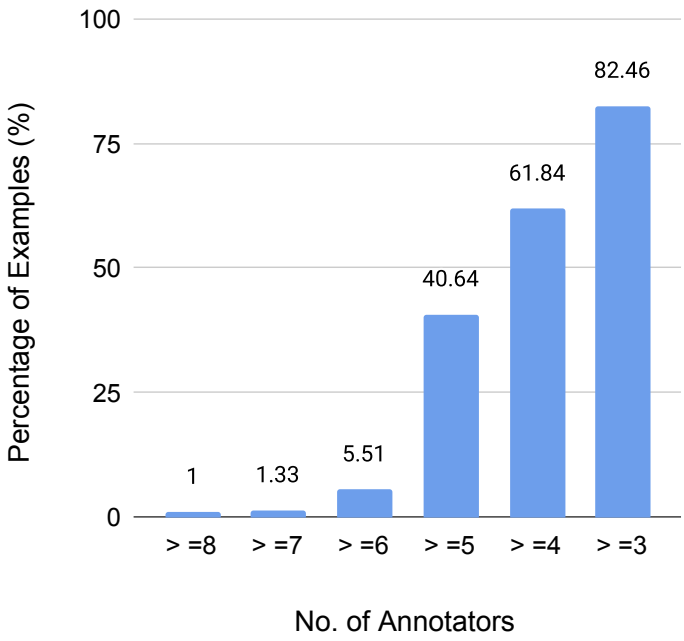


Figure C.3: Percentage of example pairs (y-axis) verses number of annotations with the exact same labeling for relevant rows (x-axis).

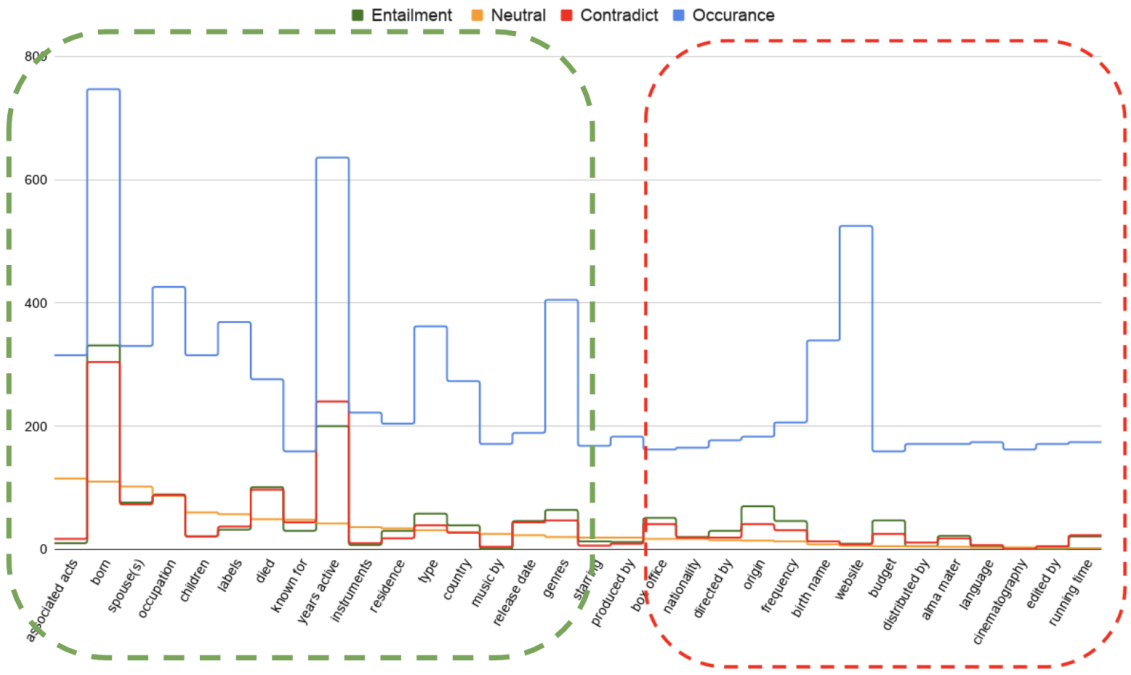


Figure C.4: Figure depicting the human bias problem with the annotations for the ENTAIL, CONTRADICT and NEUTRAL label. Here, Green and Red circles represent overused and underused keys by hypothesis, respectively.

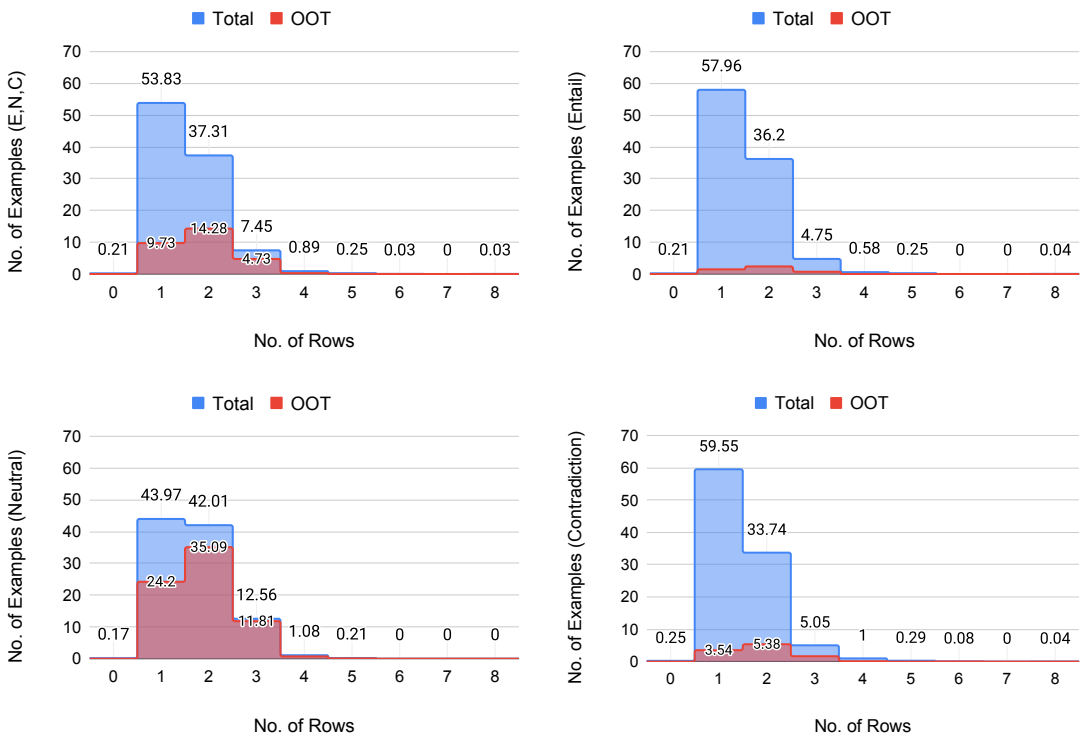


Figure C.5: Figure depicting the percentage of examples where multiple rows are relevant for a given hypothesis for ENTAIL, CONTRADICT, NEUTRAL, and in total. It also shows the percentage of examples where Out of Table(OOT) information is used.

Table C.1: Macro average of the annotators' agreement with the majority selection as the relevant rows.

Cleaning	#Anno	Prec	Recall	F1-score
Before	72	87.39	87.86	87.51
After	50	88.86	89.63	89.15

Table C.2: Percentage of annotated examples for each Fleiss' kappa bucket.

Agreement	Range	Percentage (%)
Poor	< 0	00.17
Slight	0.01 – 0.20	01.53
Fair	0.21 – 0.40	05.88
Moderate	0.41 - 0.60	14.60
Substantial	0.61 - 0.80	24.40
Perfect	0.81 - 1.00	53.43

Table C.3: Final inter annotator agreement average of all example pairs for the majority selection label w.r.t all the annotations for the pair.

Type	Precision	Recall	F1-score
macro-avg	88.8	91.0	89.0
micro-avg	88.2	89.0	88.6

APPENDIX D

TRUSTWORTHY TABULAR INFERENCE

D.1 Qualitative Examples

We manually inspect the Type I and Type II error instances for the supervised model and human annotation for the development set. Below, we show some of these examples where models conflict with ground-truth human annotation. We also provide a possible reason behind the model mistakes.

Type I: Below, we show Type I error examples.

Example I **Row**: Colorado Springs, Colorado is a poor training location for endurance athletes.

Hypothesis: The elevation of Colorado Springs, Colorado is 6,035 ft (1,839 m).

Model Prediction: Not Relevant
Human Ground Truth: Relevant Evidence.

Possible Reason: Model wasn't able to connect the concept of elevation with the perfect high elevation training ground requirement of endurance athletes. Requires common sense and knowledge.

Example II
Row: The number of number of employees of International Fund for Animal Welfare - ifaw is 300+ (worldwide).

Hypothesis: International Fund for Animal Welfare - ifaw is a national organization focused on only North America.

Model Prediction: Not Relevant
Human Ground Truth: Relevant Evidence.

Possible Reason: Model wasn't able to connect the clue ('worldwide') in the table row with the phrase 'focused on only north America'.

Example III
Row: The equipment of Combined driving are horse, carriage, horse harness equipment.

Hypothesis: Combined driving is a horse racing event style.

Model Prediction: Not Relevant
Human Ground Truth: Relevant Evidence.

Possible Reason: Model wasn't able to connect the horse related equipment i.e. 'horse carriage, horse harness' with the event time i.e. 'horse racing'.

Type II. Below, we show Type II error examples.

Example I
Row: Dazed and Confused was directed by Richard Linklater.

Hypothesis: Dazed and Confused was directed in 1993.

Model Prediction: Relevant Evidence
Human Ground Truth: Not Relevant.

Possible Reason: Model focuses on lexical match token 'directed' instead using entity type where premise refer for 'Person' who directed rather than 'Date' of direction.

Example II
Row: The spouse(s) of Celine Dion (CC OQ ChLD) is René Angélil, (m. 1994; died 2016).

Hypothesis: Thérèse Tanguay Dion had a child that became a widow.

Model Prediction: Relevant Evidence
Human Ground Truth: Not Relevant.

Possible Reason: Model was unable to connect widow concept in hypothesis with it relation to Spouse and the marriage date René Angélil, (m. 1994; died 2016).

Example III
Row: The trainer of Caveat is Woody Stephens.

Hypothesis: Caveat won more in winnings than it took to raise and train him.

Model Prediction: Relevant Evidence
Human Ground Truth: Not Relevant.

Possible Reason: Model connects the 'raise and train' term with the trainer name which is unrelated and has no connection to overall, winning races money vs spending for animal.

Discussion: Based on the observation from the above examples as also stated in 7.6.3, the model fails on many examples due to its lack of knowledge and common-sense reasoning ability. One possible solution to mitigate this is by the addition of implicit and explicit knowledge on-the-fly for evidence extraction, as done for inference task by [182].

D.2 Implicit Relevance Indication

We manually examine the human-annotated evidence in the development set. We discovered the existence of several relevant phrases/tokens which implicitly indicate the presence of evidence rows. E.g. The existence of tokens such as *married*, *husband*, *lesbian*, and *wife* in hypothesis (H) is very suggestive of the row *Spouse* being the relevant evidence. Learning such implicit relevance-based phrases and tokens connection is easy for humans and large pre-trained supervision models. It is a challenging task for similarity-based

unsupervised extraction methods. Below, we show implicit relevance, indicating token and the corresponding relevant evidence rows.

Relevance Indicating Phrase (H) → Relevant Evidence Rows Key(T)

'broked', 'started from', 'doesn't anymore', 'still perform', 'over a decade', 'began performing', 'started wrapping', 'first started' → year active
age related term, 'were of ¡age¿', 'after ¡age¿', 'fall', 'spring','birthday' → born
'several years', 'one month', century art → years
'co-wrote', 'written', 'writer', 'original written' → written by (novel and book)
'married', 'husband', 'lesbian', 'wives' → Spouse
'no-reward', 'monetary value', 'prize' → rewards
'earlier', 'debut', '21st century', 'early 90s', 'recording','product of years' → recorded
'lost', 'won', 'races','competition' → records (horse races, car races etc) 'sea level' → 'lowest elevation', 'highest elevation', 'elevation'
multi-lingual, multi-faith → 'regional languages', 'official languages', 'religion', ','race or faith'
'acting', 'rapping', 'politics' → occupation
'over an', 'shortest', 'longest', 'run-time' → length 'is form ¡country¿', 'originate', 'are an ¡nationality¿', 'formed on ¡location¿', 'moved to ¡Country¿', 'descended from' → origin, descendant, parenthood etc
'city' with 'x' peoples → 'metropolitan municipality' or 'metro'
'was painted with', 'mosaic', 'oil', 'water' → medium
'hung in' , 'museum', 'is stored in/at', 'wall', 'mural' → 'location'
'was discontinued', 'awards' → 'last awarded'
'playing bass' → 'instruments'
'served', 'term', 'current charge' , 'in-charge' → 'in office'
'is controlled by', 'under control' → 'government'
'classical', 'pop', 'rock', 'hip-hop', 'sufi' → genre
'won more', 'in winning (race)', 'earned more than' → earnings
'Register of', 'Cultural Properties' → designated
'urban area', 'less dense' -¿ urban density, density
'founded by', 'has been around', 'years' → founded , introduce
'was started', 'century', 'was formed', '100 years' → founded, formation
'daughters', 'sons' → children spouse(s), partner(s)
'lost money', 'net profit', 'budget', 'unprofitable', 'not popular'(common sense)
'owned' or 'company' → manufacturer
'bigger than an average' → dimension

APPENDIX E

TABULAR AUGMENTATION: ALBERTA PERFORMANCE

We perform a similar analysis on $\text{ALBERT}_{\text{BASE}}$ as we have done for $\text{RoBERTa}_{\text{BASE}}$ to see if our data benefits there too. To see how robust AUTO-TNLI is when improving performance in the Augmentation setting, we perform the same experiments as RQ2a in Section 8.6.3. We also explore some experiments from RQ1b in Section 8.6.2 which are shown in Table E.1.

Analysis: As we can see in Tables E.2, E.3, and E.4, the trends are very similar to what we have seen in main paper Section 8.6.3 for full supervision setting. Thus our approach of semi-automatic generation is generalizable across similar models.

Table E.1: Performance (accuracy) on AUTO-TNLI with $\text{ALBERT}_{\text{BASE}}$ model across several evaluation splits with fine-tuning on AUTO-TNLI. **bold** - represents max across rows i.e. best train/augmentation setting.

Augmentation Strategy	Cat-Ran	Cross-Cat	Key	No-Para	Cross-Para	Entity
Random	50.00	50.00	50.00	50.00	50.00	50.00
AUTO-TNLI	77.16	69.73	81.91	86.22	**87.45**	72.75
MNLI +AUTO-TNLI	**80.28**	**76.24**	**83.1**	**88.73**	87.44	**74.53**

Table E.2: Performance (accuracy) of stage one $\text{ALBERT}_{\text{BASE}}$ (i.e. NEUTRAL verses NON-NEUTRAL) across several data augmentation settings. Here, No-Augmentation means INFOTABS, and MNLI means MNLI + INFOTABS. **bold** same as Table 8.12.

Test-split	No-Augmentation	MNLI
dev	79.11	**85.22**
α_1	78.61	**82.83**
α_2	80.89	**85.22**
α_3	67.78	**73.94**

Table E.3: Accuracy of combine stage I i.e. NEUTRAL verses NON-NEUTRAL and stage II i.e. ENTAIL verses CONTRADICT classifiers ($\text{ALBERT}_{\text{BASE}}$) across several data augmentation settings. Here, for stage one we also explore pre-fine tuning on MNLI data. **bold** - represents max across columns i.e. the best augmentation setting.

Stage 2: Entail verses Contradict					
Split	No Augmentation	Orig	Orig+Count	MNLI MNLI+Orig	MNLI MNLI+Orig+Count
Stage 1: INFOTABS					
dev	60.78	61.72	62.83	**64.83**	63.89
α_1	60.89	61.33	62.78	**63.22**	63.11
α_2	49.83	53.06	51.67	55.67	**56.5**
α_3	49.39	50.11	51.72	**52.94**	51.72
Stage 1: MNLI +INFOTABS					
dev	66.28	67.44	68.22	**70.67**	69.61
α_1	65.72	66.06	67.28	67.44	**67.5**
α_2	54	56.72	55.83	60.11	**60.89**
α_3	53.33	55.11	56.11	**57.39**	56.94

Table E.4: Accuracy of stage II i.e. ENTAIL verses CONTRADICT classifiers ($\text{ALBERT}_{\text{BASE}}$) across several data augmentation settings. **bold** same as Table 8.12.

Split	No Augmentation	Orig	Orig+Count	MNLI MNLI+Orig	MNLI MNLI+Orig+Count
dev	68.92	71.25	72.33	**76**	74.5
α_1	69.42	70.92	72.92	**73.92**	73.25
α_2	47.58	52.75	50.83	58.17	**58.67**
α_3	61	64.17	66.33	**68.33**	68.08

APPENDIX F

XINFOTABS: CROSS LINGUAL TRANSFER

We are also interested in knowing whether training in one language can help transfer knowledge across other languages or not. We answer the question: *What are models of cross-lingual transfer performance?*. Since we have separate models trained on languages from our dataset available, we tested them on all other languages other than the training language to study cross-lingual transfer. The TrLangAvg scores (Training Language Average) from Tables F.1, F.2 and F.3 show how models trained on INFOTABS for one language perform on other languages for α_1, α_2 and α_3 sets respectively. XLM-R (XNLI) outperforms mBERT across all tasks. English has the best cross-lingual transferability on mBERT, whereas Spanish has the best cross-lingual transferability on XLM-R(XNLI) for the α_1 set. On mBERT, German has the best cross-lingual transferability for the α_2 dataset. On XLM-R (XNLI), German and Spanish have the best cross-lingual transferability. On mBERT, English has the best cross-lingual transferability for the α_3 dataset. On XLM-R (XNLI), English and Spanish have the best cross-lingual transferability. Furthermore, the EvLangAvg score (Evaluation Language Average) score was comparable for all languages except approximately 4% lower for Arabic ('ar') language with XLM-R(XNLI) model on all three test sets. Overall, we observe that finetuning models on high resource languages improve their cross-lingual transfer capacity considerably more than finetuning models on low resource languages.

Table F.1: Evaluation of cross lingual transfer abilities of models on α_1 evaluation set. TrLang refers to the language the model has been finetuned on and EvLang refers to the language the model has been evaluated on. **Purple**, **Orange** and **Blue** represent the highest score in the row, column and both together respectively.

Test-Split	Model	TrLang	en	de	fr	es	af	ru	zh	ar	ko	hi	TrLangAvg
α_1	$\text{mBERT}_{\text{BASE}}$	en	**67**	64	63	62	61	61	60	56	58	58	**61**
		de	63	**65**	61	62	60	59	57	56	56	57	60
		fr	64	62	**65**	62	61	59	59	55	53	57	60
		es	62	62	**63**	**63**	61	60	60	**57**	57	58	60
		af	**62**	61	61	60	**62**	59	57	55	55	55	59
		ru	63	61	61	60	59	**64**	59	56	55	55	59
		zh	55	56	58	56	59	57	**63**	55	57	58	57
		ar	57	**58**	**58**	57	**58**	**58**	57	**57**	53	57	57
		ko	58	59	58	57	57	56	58	55	**61**	57	58
		hi	59	58	59	58	57	58	58	56	54	**63**	58
	EvLangAvg		**61**	**61**	**61**	60	60	59	59	56	56	58	59
	XLM-R (XNLI)	en	**76**	73	71	73	71	**71**	71	63	70	69	71
		de	74	**75**	**74**	72	71	70	69	63	**71**	68	71
		fr	73	**74**	**74**	72	**72**	70	71	64	70	70	71
		es	**74**	73	**74**	**74**	**72**	**71**	72	65	**71**	69	**72**
		af	**72**	**72**	71	71	**72**	70	70	63	70	68	70
		ru	**73**	**73**	72	71	71	**71**	71	64	70	67	70
		zh	72	72	70	71	70	69	**73**	64	70	69	70
		ar	**71**	**71**	70	70	69	70	**71**	**68**	70	68	70
		ko	**72**	71	**72**	71	70	69	71	64	**71**	69	70
		hi	**73**	**73**	71	72	70	70	70	64	69	**71**	70
	EvLangAvg		**73**	**73**	72	72	71	70	71	64	70	69	70

Table F.2: Evaluation of cross lingual transfer abilities of models on α_2 evaluation set. TrLang refers to the language the model has been finetuned on and EvLang refers to the language the model has been evaluated on. **Purple**, **Orange** and **Blue** represent the highest score in the row, column and both together respectively.

Test-Split	Model	TrLang	en	de	fr	es	af	ru	zh	ar	ko	hi	TrLangAvg
α_2	$\text{mBERT}_{\text{BASE}}$	en	**54**	53	53	53	51	52	50	49	50	47	51
		de	**54**	**54**	53	53	52	52	50	49	50	48	52
		fr	52	51	52	**53**	50	50	48	49	51	47	50
		es	52	50	50	**53**	47	51	48	49	46	46	49
		af	49	**50**	**50**	49	**50**	**50**	47	48	48	46	49
		ru	51	50	51	51	51	**52**	49	49	49	49	50
		zh	49	48	49	48	49	49	**52**	47	48	48	49
		ar	**49**	48	**49**	48	47	48	47	48	47	47	48
		ko	49	49	50	48	48	47	50	47	**51**	49	49
		hi	48	47	47	48	48	49	48	46	48	**50**	48
		EvLangAvg	**51**	50	50	50	49	50	49	48	49	48	50
	XLM-R (XNLI)	en	**68**	65	64	64	**64**	63	62	**58**	63	59	63
		de	**67**	**66**	**66**	65	**64**	63	62	57	**64**	61	**64**
		fr	**67**	64	64	65	62	60	60	**58**	62	60	62
		es	**67**	**66**	65	**66**	63	**64**	62	57	**64**	61	64
		af	**66**	64	64	64	63	62	63	57	62	59	62
		ru	**66**	64	64	63	62	**64**	62	57	61	60	62
		zh	**67**	65	65	64	63	**64**	**64**	**58**	**64**	61	62
		ar	**64**	61	62	61	60	60	60	57	60	58	60
		ko	**65**	63	63	63	61	62	62	57	**64**	59	62
		hi	**67**	64	65	65	63	**64**	62	**58**	60	**62**	63
		EvLangAvg	**66**	64	64	64	63	63	62	57	62	60	63

Table F.3: Evaluation of cross lingual transfer abilities of models on α_3 evaluation set. TrLang refers to the language the model has been finetuned on and EvLang refers to the language the model has been evaluated on. **Purple**, **Orange** and **Blue** represent the highest score in the row, column and both together respectively.

Test-Split	Model	TrLang	en	de	fr	es	af	ru	zh	ar	ko	hi	TrLangAvg
α_3	$\text{mBERT}_{\text{BASE}}$	en	**52**	**52**	51	**53**	49	**50**	49	47	46	47	**50**
		de	50	50	**51**	50	**51**	48	48	44	46	48	49
		fr	**52**	**52**	**52**	**53**	50	**50**	49	46	44	47	**50**
		es	50	50	51	**53**	48	48	46	46	46	46	**50**
		af	50	50	50	**51**	50	49	47	47	45	48	49
		ru	**50**	48	49	**50**	49	**50**	47	45	45	46	48
		zh	49	49	50	50	49	**50**	**51**	46	**48**	49	49
		ar	**49**	**49**	**49**	**49**	48	**49**	48	**49**	47	48	48
		ko	47	46	47	47	44	45	45	43	**48**	**48**	46
		hi	**50**	49	49	49	48	46	48	46	47	**50**	48
		EvLangAvg	**50**	49	**50**	**50**	49	48	48	46	46	48	49
	XLM-R (XNLI)	en	**67**	**65**	61	**64**	62	**64**	**63**	58	**65**	**62**	**63**
		de	**65**	**65**	**63**	61	**63**	63	61	56	61	60	62
		fr	**66**	64	62	63	62	61	61	56	60	**62**	62
		es	**66**	**65**	**63**	**64**	**63**	63	62	**59**	61	**62**	**63**
		af	**65**	64	61	62	62	60	61	56	60	59	61
		ru	**65**	63	61	62	62	62	61	56	60	**62**	61
		zh	**65**	64	62	63	62	62	**63**	57	62	60	62
		ar	**63**	62	62	61	61	60	60	57	60	60	61
		ko	**64**	62	61	62	60	63	61	56	60	**62**	61
		hi	**64**	63	62	63	61	61	60	58	60	**62**	61
		EvLangAvg	**65**	64	62	63	62	62	61	57	61	61	62

REFERENCES

[1] F. Abbas, M. K. Malik, M. Rashid, and R. Zafar, *WikiQA — A question answering system on Wikipedia using freebase, DBpedia and Infobox*, 2016 Sixth International Conference on Innovative Computing Technology (INTECH), IEEE, 2016, pp. 185–193.

[2] K. Acharya, *KaushikAcharya at SemEval-2021 task 9: Candidate generation for fact verification over tables*, in Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, Aug. 2021, Association for Computational Linguistics, pp. 1271–1275.

[3] A. F. Agarap, *Deep Learning Using Rectified Linear Units (ReLU)*, preprint, arXiv, 2018.

[4] C. Agarwal, V. Gupta, A. Kunchukuttan, and M. Shrivastava, *Bilingual tabular inference: A case study on Indic languages*, in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, July 2022, Association for Computational Linguistics, pp. 4018–4037.

[5] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal, *The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task*, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 1–13.

[6] J. Andreas, *Good-enough compositional data augmentation*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 7556–7566.

[7] D. Angelov, *Top2vec: Distributed Representations of Topics*, preprint, arXiv, 2020.

[8] M. Artetxe, S. Ruder, and D. Yogatama, *On the cross-lingual transferability of monolingual representations*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4623–4637.

[9] R. Artstein and M. Poesio, *Inter-coder agreement for computational linguistics*, Comput. Linguist., 34 (2008), pp. 555–596.

[10] S. Asaadi, S. Mohammad, and S. Kiritchenko, *Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 505–516.

[11] J. A. Bateman, *Towards meaning-based machine translation: Using abstractions from text generation for preserving meaning*, Mach. Trans., 7 (1992), pp. 5–40.

[12] L. Bauer, L. Deng, and M. Bansal, *ERNIE-NLI: Analyzing the impact of domain-specific external knowledge on enhanced representations for NLI*, in Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Online, June 2021, Association for Computational Linguistics, pp. 58–69.

[13] I. Beltagy, M. E. Peters, and A. Cohan, *Longformer: The Long-Document Transformer*, preprint, arXiv, 2020.

[14] C. Bhagavatula, R. L. Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, S. W.-t. Yih, and Y. Choi, *Abductive commonsense reasoning*, in International Conference on Learning Representations, Online, 2020, ICLR.

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, *Enriching word vectors with subword information*, Trans. Assoc. Comput. Linguist., 5 (2017), pp. 135–146.

[16] M. Bouziane, H. Perrin, A. Sadeq, T. Nguyen, A. Cluzeau, and J. Mardas, *FaBULOUS: Fact-checking based on understanding of language over unstructured and structured information*, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 31–39.

[17] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, *A large annotated corpus for learning natural language inference*, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Sept. 2015, Association for Computational Linguistics, pp. 632–642.

[18] M. Bronzi, V. Crescenzi, P. Merialdo, and P. Papotti, *Extraction and integration of partially overlapping web sources*, vol. 6, Riva del Garda, Italy, August 2013, VLDB Endowment, p. 805–816.

[19] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, *SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation*, in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada, Aug. 2017, Association for Computational Linguistics, pp. 1–14.

[20] V. Cetorelli, P. Atzeni, V. Crescenzi, and F. Milicchio, *The smallest extraction problem*, in Proceedings of the VLDB Endowment, Copenhagen, Denmark, 2021, VLDB Endowment.

[21] T.-Y. Chang, Y. Liu, K. Gopalakrishnan, B. Hedayatnia, P. Zhou, and D. Hakkani-Tur, *Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks*, in Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Online, Nov. 2020, Association for Computational Linguistics, pp. 74–79.

[22] M. Chen, H. Zhang, Q. Ning, M. Li, H. Ji, K. McKeown, and D. Roth, *Event-centric natural language processing*, in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts, Online, Aug. 2021, Association for Computational Linguistics, pp. 6–14.

[23] Q. Chen, X. Zhu, Z.-H. Ling, D. Inkpen, and S. Wei, *Neural natural language inference models enhanced with external knowledge*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 2018, Association for Computational Linguistics, pp. 2406–2417.

[24] W. Chen, M.-W. Chang, E. Schlinger, W. Wang, and W. W. Cohen, *Open question answering over tables and text*, in International Conference on Learning Representations, Online, 2020, ICLR.

[25] W. Chen, J. Chen, Y. Su, Z. Chen, and W. Y. Wang, *Logical natural language generation from open-domain tables*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 7929–7942.

[26] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang, *TabFact: A Large-scale Dataset for Table-based Fact Verification*, in International Conference on Learning Representations, Addis Ababa, Ethiopia, April 2020, ICLR.

[27] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Y. Wang, *HybridQA: A dataset of multi-hop question answering over tabular and textual data*, in Findings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov. 2020, Association for Computational Linguistics, pp. 1026–1036.

[28] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, and H. Chen, *Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction*, in Proceedings of the ACM Web Conference 2022, WWW '22, New York, NY, USA, 2022, Association for Computing Machinery, p. 2778–2788.

[29] Z. Chen, W. Chen, H. Zha, X. Zhou, Y. Zhang, S. Sundaresan, and W. Y. Wang, *Logic2Text: High-fidelity natural language generation from logical forms*, in Findings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov. 2020, Association for Computational Linguistics, pp. 2096–2111.

[30] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah, and W. Y. Wang, *ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering*, in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, Dec. 2022, Association for Computational Linguistics, pp. 6279–6292.

[31] Z. Cheng, H. Dong, Z. Wang, R. Jia, J. Guo, Y. Gao, S. Han, J.-G. Lou, and D. Zhang, *HiTab: A hierarchical table dataset for question answering and natural language generation*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 1094–1110.

[32] Z. Chi, S. Huang, L. Dong, S. Ma, B. Zheng, S. Singhal, P. Bajaj, X. Song, X.-L. Mao, H. Huang, and F. Wei, *XLM-E: Cross-lingual language model pre-training via ELECTRA*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 6170–6182.

[33] T. Chiang, *On a Benefit of Mask Language Modeling: Robustness to Simplicity Bias*, preprint, arXiv, 2021.

[34] D. Chicco, *Siamese neural networks: An overview*, in Artificial Neural Networks, H. Cartwright, ed., 2021, Springer US, New York, pp. 73–94.

[35] H. W. Chung, T. Fevry, H. Tsai, M. Johnson, and S. Ruder, *Rethinking embedding coupling in pre-trained language models*, in International Conference on Learning Representations, ICLR, 2021.

[36] F. Ciravegna, A. L. Gentile, and Z. Zhang, *LODIE: Linked open data for web-scale information extraction*, SWAIE, 925 (2012), pp. 11–22.

[37] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki, *TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages*, Trans. Assoc. Comput. Linguist., 8 (2020), pp. 454–470.

[38] P. Clark and O. Etzioni, *My computer is an honor student — But how intelligent is it? Standardized tests as a measure of AI*, AI Mag., 37 (2016), pp. 5–12.

[39] R. Cohn-Gordon and N. Goodman, *Lost in machine translation: A method to reduce meaning loss*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 437–441.

[40] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, *Unsupervised cross-lingual representation learning at scale*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 8440–8451.

[41] A. Conneau and G. Lample, *Cross-lingual language model pretraining*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019, pp. 7059–7069.

[42] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, *XNLI: Evaluating cross-lingual sentence representations*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Oct.–Nov. 2018, Association for Computational Linguistics, pp. 2475–2485.

[43] V. Crescenzi, G. Mecca, P. Merialdo, et al., *Roadrunner: Towards automatic data extraction from large web sites*, in VLDB, vol. 1, VLDB Endowment, 2001, pp. 109–118.

[44] I. Dagan, O. Glickman, and B. Magnini, *The Pascal recognising textual entailment challenge*, in Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment, MLCW'05, Berlin, Heidelberg, 2005, Springer-Verlag, p. 177–190.

[45] I. Dagan, D. Roth, M. Sammons, and F. M. Zanzotto, *Recognizing textual entailment: Models and applications*, Synth. Lect. Hum. Lang. Technol., 6 (2013), pp. 1–220.

[46] B. Dalvi Mishra, N. Tandon, and P. Clark, *Domain-targeted, high precision knowledge extraction*, Trans. Assoc. Comput. Linguist., 5 (2017), pp. 233–246.

[47] D. Demszky, K. Guu, and P. Liang, *Transforming Question Answering Datasets into Natural Language Inference Datasets*, preprint, arXiv, 2018.

[48] D. Deng, Y. Jiang, G. Li, J. Li, and C. Yu, *Scalable column concept determination for web tables using large knowledge bases*, Proc. VLDB Endow., 6 (2013), p. 1606–1617.

[49] X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, *TURL: Table understanding through representation learning*, in SIGMOD Rec., vol. 51, New York, NY, USA, Jun. 2022, Association for Computing Machinery, p. 33–40.

[50] A. Deshpande, P. Talukdar, and K. Narasimhan, *When is BERT multilingual? Isolating crucial ingredients for cross-lingual transfer*, in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, July 2022, Association for Computational Linguistics, pp. 3610–3623.

[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 4171–4186.

[52] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, *ERASER: A benchmark to evaluate rationalized NLP models*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4443–4458.

[53] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, and W. W. Cohen, *Time-aware language models as temporal knowledge bases*, Trans. Assoc. Comput. Linguist., 10 (2022), pp. 257–273.

[54] B. Dhingra, M. Faruqui, A. Parikh, M.-W. Chang, D. Das, and W. Cohen, *Handling divergent reference texts when evaluating table-to-text generation*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 4884–4895.

[55] X. L. Dong, H. Hajishirzi, C. Lockard, and P. Shiralkar, *Multi-modal information extraction from text, semi-structured, and tabular data on the web*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Online, July 2020, Association for Computational Linguistics, pp. 23–26.

[56] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res., 12 (2011), pp. 2121–2159.

[57] J. Dwivedi-Yu, T. Schick, Z. Jiang, M. Lomeli, P. Lewis, G. Izacard, E. Grave, S. Riedel, and F. Petroni, *EditEval: An Instruction-Based Benchmark for Text Improvements*, preprint, arXiv, 2022..

[58] J. Eisenschlos, S. Krichene, and T. Müller, *Understanding tables with intermediate pre-training*, in Findings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov. 2020, Association for Computational Linguistics, pp. 281–296.

[59] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al., *Beyond English-centric multilingual machine translation*, J. Mach. Learn. Res., 22 (2021), pp. 1–48.

[60] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, *Beyond English-Centric Multilingual Machine Translation*, preprint, arXiv, 2020.

[61] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, *Beyond English-Centric Multilingual Machine Translation*, 2020.

[62] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, *LIBLINEAR: A library for large linear classification*, J. Mach. Learn. Res., 9 (2008), p. 1871–1874.

[63] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber, *Pathologies of neural models make interpretations difficult*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Oct.–Nov. 2018, Association for Computational Linguistics, pp. 3719–3728.

[64] E. Fonseca, L. Yankovskaya, A. F. T. Martins, M. Fishel, and C. Federmann, *Findings of the WMT 2019 shared tasks on quality estimation*, in Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), Florence, Italy, Aug. 2019, Association for Computational Linguistics, pp. 1–10.

[65] M. Funkquist, *Combining sentence and table evidence to predict veracity of factual claims using TaPaS and RoBERTa*, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 92–100.

[66] T. Furche, G. Gottlob, G. Grasso, O. Gunes, X. Guo, A. Kravchenko, G. Orsi, C. Schallhart, A. Sellers, and C. Wang, *DIADEM: Domain-centric, intelligent, automated data extraction methodology*, in Proceedings of the 21st International Conference on World Wide Web, WWW '12 Companion, Lyon, France, 2012, Association for Computing Machinery, p. 267–270.

[67] T. Furche, G. Gottlob, G. Grasso, X. Guo, G. Orsi, C. Schallhart, and C. Wang, *DIADEM: thousands of websites to a single database*, Proc. VLDB Endowment, 7 (2014), pp. 1845–1856.

[68] A. Gajbhiye, N. A. Moubayed, and S. Bradley, *ExBERT: An external knowledge enhanced BERT for natural language inference*, in Artificial Neural Networks and Machine Learning – ICANN 2021, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, eds., Cham, 2021, Springer International Publishing, pp. 460–472.

[69] A. Gajbhiye, T. Winterbottom, N. Al Moubayed, and S. Bradley, *Bilinear fusion of commonsense knowledge with attention-based NLI models*, in International Conference on Artificial Neural Networks, Springer, 2020, pp. 633–646.

[70] Y. Gal and Z. Ghahramani, *A theoretically grounded application of dropout in recurrent neural networks*, in Advances in Neural Information Processing Systems, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., vol. 29, Curran Associates, Inc., 2016.

[71] T. Gao, X. Yao, and D. Chen, *SimCSE: Simple contrastive learning of sentence embeddings*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 6894–6910.

[72] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, N. Gupta, H. Hajishirzi, G. Ilharco, D. Khashabi, K. Lin, J. Liu, N. F. Liu, P. Mulcaire, Q. Ning, S. Singh, N. A. Smith, S. Subramanian, R. Tsarfaty, E. Wallace, A. Zhang, and B. Zhou, *Evaluating models' local decision boundaries via contrast sets*, in Findings of the Association for Computational Linguistics, 2020, pp. 1307–1323.

[73] P. Gatti, A. Mishra, M. Gupta, and M. Das Gupta, *VisToT: Vision-augmented table-to-text generation*, in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, Dec. 2022, Association for Computational Linguistics, pp. 9936–9949.

[74] D. Gautam, K. Gupta, and M. Shrivastava, *Volta at SemEval-2021 task 9: Statement verification and evidence finding with tables using TAPAS and transfer learning*, in Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, Aug. 2021, Association for Computational Linguistics, pp. 1262–1270.

[75] M. Geva, Y. Goldberg, and J. Berant, *Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 1161–1166.

[76] I.-Z. Gi, T.-Y. Fang, and R. T.-H. Tsai, *Verdict inference with claim and retrieved elements using RoBERTa*, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 60–65.

[77] M. Glass, M. Canim, A. Gliozzo, S. Chemmengath, V. Kumar, R. Chakravarti, A. Sil, F. Pan, S. Bharadwaj, and N. R. Fauceglia, *Capturing row and column semantics in transformer based question answering over tables*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 1212–1224.

[78] M. Glockner, V. Shwartz, and Y. Goldberg, *Breaking NLI systems with sentences that require simple lexical inferences*, in Proceedings of the 56th Annual Meeting of the Asso-

ciation for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, July 2018, Association for Computational Linguistics, pp. 650–655.

[79] K. Goel, N. F. Rajani, J. Vig, Z. Taschdjian, M. Bansal, and C. Ré, *Robustness gym: Unifying the NLP evaluation landscape*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, Online, June 2021, Association for Computational Linguistics, pp. 42–55.

[80] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, *An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks*, preprint, arXiv, 2015.

[81] P. Gulhane, A. Madaan, R. Mehta, J. Ramamirtham, R. Rastogi, S. Satpal, S. H. Sengamedu, A. Tengli, and C. Tiwari, *Web-scale information extraction with vertex*, in 2011 IEEE 27th International Conference on Data Engineering, IEEE, 2011, pp. 1209–1220.

[82] R. Gupta, A. Halevy, X. Wang, S. E. Whang, and F. Wu, *Biperpedia: An ontology for search applications*, Proc. VLDB Endow., 7 (2014), p. 505–516.

[83] V. Gupta, R. A. Bhat, A. Ghosal, M. Shrivastava, M. Singh, and V. Srikumar, *Is my model using the right evidence? Systematic probes for examining evidence-based tabular reasoning*, Trans. Assoc. Comput. Linguist., 10 (2022), pp. 659–679.

[84] V. Gupta, M. Mehta, P. Nokhiz, and V. Srikumar, *INFOTABS: Inference on tables as semi-structured data*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 2309–2324.

[85] V. Gupta, S. Zhang, A. Vempala, Y. He, T. Choji, and V. Srikumar, *Right for the right reason: Evidence extraction for trustworthy tabular reasoning*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 3268–3283.

[86] I. Gurevych and N. Reimers, *Making monolingual sentence embeddings multilingual using knowledge distillation*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 4512–4525.

[87] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, *Annotation artifacts in natural language inference data*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 107–112.

[88] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, *Annotation artifacts in natural language inference data*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 107–112.

[89] E. Haihong, W. Zhang, and M. Song, *KB-Transformer: Incorporating knowledge into end-to-end task-oriented dialog systems*, in 2019 15th International Conference on Semantics, Knowledge and Grids (SKG), IEEE, 2019, pp. 44–48.

[90] Q. Hao, R. Cai, Y. Pang, and L. Zhang, *From one tree to a forest: A unified solution for structured web data extraction*, in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, New York, NY, USA, 2011, Association for Computing Machinery, p. 775–784.

[91] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[92] P. He, X. Liu, J. Gao, and W. Chen, *DeBERTa: Decoding-enhanced BERT with disentangled attention*, in International Conference on Learning Representations, Online, 2021, ICLR Openreview.

[93] J. Herzig, T. Müller, S. Krichene, and J. Eisenschlos, *Open domain question answering over tables via dense retrieval*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 512–519.

[94] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, and J. Eisenschlos, *TaPas: Weakly supervised table parsing via pre-training*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4320–4333.

[95] J. Hewitt and P. Liang, *Designing and interpreting probes with control tasks*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 2733–2743.

[96] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural Comput., 9 (1997), pp. 1735–1780.

[97] Y. Hou, G. Fu, and M. Sachan, *Understanding Knowledge Integration in Language Models with Graph Convolutions*, preprint, arXiv, 2022.

[98] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, *XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation*, in International Conference on Machine Learning, PMLR, 2020, pp. 4411–4421.

[99] W. Huang, H. Liu, and S. Bowman, *Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data*, in Proceedings of the First Workshop on Insights from Negative Results in NLP, Online, Nov. 2020, Association for Computational Linguistics, pp. 82–87.

[100] M. Hulsebos, K. Hu, M. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, and C. Hidalgo, *Sherlock: A deep learning approach to semantic data type detection*, in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining, KDD '19, Anchorage, AK, USA, 2019, Association for Computing Machinery, p. 1500–1508.

[101] H. Iida, D. Thai, V. Manjunatha, and M. Iyyer, *TABBIE: Pretrained representations of tabular data*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 3446–3456.

[102] R. Iv, A. Passos, S. Singh, and M.-W. Chang, *FRUIT: Faithfully reflecting updated information in text*, in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, July 2022, Association for Computational Linguistics, pp. 3670–3686.

[103] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, *Adversarial example generation with syntactically controlled paraphrase networks*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 1875–1885.

[104] N. Jain, V. Gupta, A. Rai, and G. Kumar, *TabPert: An effective platform for tabular perturbation*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online and Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 350–360.

[105] S. Jain and B. C. Wallace, *Attention is not explanation*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 3543–3556.

[106] M. Jalili Sabet, P. Dufter, F. Yvon, and H. Schütze, *SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings*, in Findings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov. 2020, Association for Computational Linguistics, pp. 1627–1643.

[107] A. Jena, V. Gupta, M. Shrivastava, and J. Eisenschlos, *Leveraging data recasting to enhance tabular reasoning*, in Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, Dec. 2022, Association for Computational Linguistics, pp. 4483–4496.

[108] R. Jia and P. Liang, *Adversarial examples for evaluating reading comprehension systems*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, Sept. 2017, Association for Computational Linguistics, pp. 2021–2031.

[109] Z. Jia, A. Abujabal, R. Saha Roy, J. Strötgen, and G. Weikum, *TempQuestions: A benchmark for temporal question answering*, in Companion Proceedings of the The Web Conference 2018, WWW '18, Lyon, France, 2018, International World Wide Web Conferences Steering Committee, pp. 1057–1062.

[110] R. Jiang, R. E. Banchs, and H. Li, *Evaluating and combining name entity recognition systems*, in Proceedings of the Sixth Named Entity Workshop, Berlin, Germany, Aug. 2016, Association for Computational Linguistics, pp. 21–27.

[111] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, *How can we know what language models know?* Trans. Assoc. Comput. Linguist., 8 (2020), pp. 423–438.

[112] X. Jin and J. Han, *K-means clustering*, in Encyclopedia of Machine Learning and Data Mining, Springer, Boston, 2010, pp. 563–564.

[113] A. Jindal, A. Gupta, J. Srivastava, P. Menghwani, V. Malik, V. Kaushik, and A. Modi, *BreakingBERT@IITK at SemEval-2021 task 9: Statement verification and evidence finding with tables*, in Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, Aug. 2021, Association for Computational Linguistics, pp. 327–337.

[114] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, *SpanBERT: Improving pre-training by representing and predicting spans*, Trans. Assoc. Comput. Linguist., 8 (2020), pp. 64–77.

[115] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. J'egou, and T. Mikolov, *FastText.zip: Compressing Text Classification Models*, preprint, arXiv, 2016.

[116] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, *Marian: Fast neural machine translation in C++*, in Proceedings of ACL 2018, System Demonstrations, Melbourne, Australia, July 2018, Association for Computational Linguistics, pp. 116–121.

[117] K. K, A. Sathe, S. Aditya, and M. Choudhury, *Analyzing the effects of reasoning types on cross-lingual transfer performance*, in Proceedings of the 1st Workshop on Multilingual Representation Learning, Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 86–95.

[118] D. Kang, T. Khot, A. Sabharwal, and E. Hovy, *AdvEntuRe: Adversarial training for textual entailment with knowledge-guided examples*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 2018, Association for Computational Linguistics, pp. 2418–2428.

[119] D. Kaushik, E. Hovy, and Z. Lipton, *Learning the difference that makes a difference with counterfactually-augmented data*, in International Conference on Learning Representations, Online, 2019, Openreview.

[120] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, *Generalization through memorization: Nearest neighbor language models*, in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

[121] S. Khanuja, M. Johnson, and P. Talukdar, *MergeDistill: Merging language models using pre-trained distillation*, in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, Aug. 2021, Association for Computational Linguistics, pp. 2874–2887.

[122] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, *Looking beyond the surface: A challenge set for reading comprehension over multiple sentences*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Compu-tational Linguistics: Human Language Technologies, 2018.

[123] D. Khashabi, T. Khot, A. Sabharwal, P. Clark, O. Etzioni, and D. Roth, *Question answering via integer programming over semi-structured knowledge*, in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, ICLR.

[124] T. Khot, A. Sabharwal, and P. Clark, *SciTaiL: A textual entailment dataset from science question answering*, AAAI, 32 (2018), pp. 5189–5197.

[125] G. Koch, R. Zemel, R. Salakhutdinov, et al., *Siamese neural networks for one-shot image recognition*, in ICML Deep Learning Workshop, vol. 2, Lille, 2015, ICML.

[126] P. Koehn and C. Monz, *Manual and automatic evaluation of machine translation between European languages*, in Proceedings on the Workshop on Statistical Machine Translation, New York City, June 2006, Association for Computational Linguistics, pp. 102–121.

[127] N. Kotonya, T. Spooner, D. Magazzeni, and F. Toni, *Graph reasoning with context-aware linearization for interpretable fact extraction and verification*, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 21–30.

[128] J. Krishnamurthy, P. Dasigi, and M. Gardner, *Neural semantic parsing with type constraints for semi-structured tables*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, Sept. 2017, Association for Computational Linguistics, pp. 1516–1526.

[129] D. Kumar, V. Gupta, S. Sharma, and S. Zhang, *Realistic data augmentation framework for enhancing tabular reasoning*, in Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, Dec. 2022, Association for Computational Linguistics, pp. 4411–4429.

[130] S. Kumar and P. Talukdar, *NILE : Natural language inference with faithful natural language explanations*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 8730–8742.

[131] N. Kushmerick, *Wrapper Induction for Information Extraction*, PhD dissertation, University of Washington, Seattle, 1997.

[132] N. Kushmerick, *Wrapper induction: Efficiency and expressiveness*, Artif. Intell., 118 (2000), pp. 15–68.

[133] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, *From word embeddings to document distances*, in Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, p. 957–966.

[134] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, *ALBERT: A lite BERT for self-supervised learning of language representations*, in International Conference on Learning Representations, ICLR, 2020.

[135] K. Lee, M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova, *Pix2struct: Screenshot Parsing as Pretraining for Visual Language Understanding*, preprint, arXiv, 2022.

[136] H. Levesque, E. Davis, and L. Morgenstern, *The Winograd Schema Challenge*, in Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning, oronto, Ontario, Canada, 2012, ICLR.

[137] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 7871–7880.

[138] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, *MLQA: Evaluating cross-lingual extractive question answering*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 7315–7330.

[139] P. Lewis, P. Stenetorp, and S. Riedel, *Question and answer test-train overlap in open-domain question answering datasets*, in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, Apr. 2021, Association for Computational Linguistics, pp. 1000–1008.

[140] A. H. Li and A. Sethy, *Knowledge Enhanced Attention for Robust Natural Language Inference*, preprint, arXiv, 2019.

[141] T. Li, L. Fang, J.-G. Lou, and Z. Li, *TWT: Table with written text for controlled data-to-text generation*, in Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 1244–1254.

[142] T. Li, P. A. Jawale, M. Palmer, and V. Srikumar, *Structured tuning for semantic role labeling*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 8402–8412.

[143] T. Li, X. Zhu, Q. Liu, Q. Chen, Z. Chen, and S. Wei, *Several Experiments on Investigating Pretraining and Knowledge-Enhanced Models for Natural Language Inference*, preprint, arXiv, 2019.

[144] X. Li and F. Orabona, *On the convergence of stochastic gradient descent with adaptive stepsizes*, in 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 983–992.

[145] Y. Li, W. Li, and L. Nie, *MMCoQA: Conversational question answering over text, tables, and images*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 4220–4231.

[146] Y. Liang, N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao, et al., *XGLUE: A New Benchmark Dataset for Cross-Lingual Pre-Training, Understanding and Generation*, preprint, arXiv, 2020.

[147] Y. Liang, N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao, X. Fan, R. Zhang, R. Agrawal, E. Cui, S. Wei, T. Bharti, Y. Qiao, J.-H. Chen, W. Wu,

S. Liu, F. Yang, D. Campos, R. Majumder, and M. Zhou, *XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 6008–6018.

[148] G. Limaye, S. Sarawagi, and S. Chakrabarti, *Annotating and searching web tables using entities, types and relationships*, Proc. VLDB Endow., 3 (2010), pp. 1338–1347.

[149] B. Y. Lin, X. Chen, J. Chen, and X. Ren, *KagNet: Knowledge-aware graph networks for commonsense reasoning*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 2829–2839.

[150] H. Lin, L. Sun, and X. Han, *Reasoning with heterogeneous knowledge for commonsense machine comprehension*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, Sept. 2017, Association for Computational Linguistics, pp. 2032–2043.

[151] X. V. Lin, R. Socher, and C. Xiong, *Bridging textual and tabular data for cross-domain Text-to-SQL semantic parsing*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Online, Nov. 2020, Association for Compu-tational Linguistics, pp. 4870–4888.

[152] F. Liu, J. M. Eisenschlos, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, W. Chen, N. Collier, and Y. Altun, *DePlot: One-Shot Visual Language Reasoning by Plot-to-Table Translation*, preprint, arXiv, 2022.

[153] F. Liu, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, Y. Altun, N. Collier, and J. M. Eisenschlos, *MatCha: Enhancing Visual Language Pretraining with Math reasoning and Chart Derendering*, preprint, arXiv, 2022.

[154] L. Z. Liu, Y. Wang, J. Kasai, H. Hajishirzi, and N. A. Smith, *Probing Across Time: What Does RoBERTa Know and When?* preprint, arXiv, 2021.

[155] N. F. Liu, R. Schwartz, and N. A. Smith, *Inoculation by fine-tuning: A method for analyzing challenge datasets*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 2171–2179.

[156] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, *Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, preprint, CoRR, 2021.

[157] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, *Multilingual denoising pre-training for neural machine translation*, Trans. Assoc. Comput. Linguist., 8 (2020), pp. 726–742.

[158] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, preprint, arXiv, 2019.

[159] C. Lockard, X. L. Dong, A. Einolghozati, and P. Shiralkar, *CERES: Distantly supervised relation extraction from the semi-structured web*, Proc. VLDB Endow., 11 (2018), pp. 1084–1096.

[160] C. Lockard, P. Shiralkar, and X. L. Dong, *OpenCeres: When open information extraction meets the semi-structured web*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 3047–3056.

[161] C. Lockard, P. Shiralkar, X. L. Dong, and H. Hajishirzi, *ZeroShotCeres: Zero-shot relation extraction from semi-structured webpages*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8105–8117.

[162] R. Logan, N. F. Liu, M. E. Peters, M. Gardner, and S. Singh, *Barack's wife Hillary: Using knowledge graphs for fact-aware language modeling*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 5962–5971.

[163] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, in International Conference on Learning Representations, Toulon, France, April 2017, ICLR.

[164] Y. Lu, H. Lu, G. Fu, and Q. Liu, *KELM: Knowledge Enhanced Pre-Trained Language Representations with Message Passing on Hierarchical Relational Graphs*, preprint, arXiv, 2021.

[165] C. Malon, *Team Papelo at FEVEROUS: Multi-hop evidence pursuit*, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 40–49.

[166] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. Jawahar, *InfographicVQA*, in 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE Computer Society, 2022, pp. 2582–2591.

[167] T. McCoy, E. Pavlick, and T. Linzen, *Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 3428–3448.

[168] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, *Advances in pre-training distributed word representations*, in Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018, European Language Resources Association (ELRA).

[169] G. Miller, *Wordnet: A lexical database for English*, Commun. ACM, 38 (1995), p. 39–41.

[170] G. A. Miller, *WordNet: A lexical database for English*, in Speech and Natural Language: Proceedings of a Workshop, Harriman, New York, February 23-26, 1992, 1992.

[171] B. Minhas, A. Shankhdhar, V. Gupta, D. Aggarwal, and S. Zhang, *XInfoTabS: Evaluating multilingual tabular natural language inference*, in Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 59–77.

[172] A. Mishra, D. Patel, A. Vijayakumar, X. L. Li, P. Kapanipathi, and K. Talamadupula, *Looking beyond sentence-level natural language inference for question answering and text summarization*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 1322–1336.

[173] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, *Cross-task generalization via natural language crowdsourcing instructions*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 3470–3487.

[174] A. Mitra, P. Banerjee, K. K. Pal, S. Mishra, and C. Baral, *How Additional Knowledge Can Improve Natural Language Commonsense Question Answering?* preprint, arXiv, 2019.

[175] M. Moradi and M. Samwald, *Evaluating the robustness of neural language models to input perturbations*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 1558–1570.

[176] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, *TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Oct. 2020, Association for Computational Linguistics, pp. 119–126.

[177] T. Müller, J. Eisenschlos, and S. Krichene, *TAPAS at SemEval-2021 task 9: Reasoning over tables with intermediate pre-training*, in Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, Aug. 2021, Association for Computational Linguistics, pp. 423–430.

[178] A. Naik, A. Ravichander, C. Rose, and E. Hovy, *Exploring numeracy in word embeddings*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 3374–3380.

[179] A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neubig, *Stress test evaluation for natural language inference*, in Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, Aug. 2018, Association for Computational Linguistics, pp. 2340–2353.

[180] K. Nakamura, S. Levy, Y.-L. Tuan, W. Chen, and W. Y. Wang, *HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data*, in Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 481–492.

[181] S. Neelam, U. Sharma, H. Karanam, S. Ikbal, P. Kapanipathi, I. Abdelaziz, N. Mihindukulasooriya, Y.-S. Lee, S. Srivastava, C. Pendus, et al., *A Benchmark for Generalizable and Interpretable Temporal Question Answering over Knowledge Bases*, preprint, arXiv, 2022.

[182] J. Neeraja, V. Gupta, and V. Srikumar, *Incorporating external knowledge to enhance tabular reasoning*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 2799–2809.

[183] Y. Nie, H. Chen, and M. Bansal, *Combining fact extraction and verification with neural semantic matching networks*, in Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19, Honolulu, Hawaii, USA, 2019, AAAI Press.

[184] Y. Nie, Y. Wang, and M. Bansal, *Analyzing compositionality-sensitivity of NLI models*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, Jul. 2019, pp. 6867–6874.

[185] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, *Adversarial NLI: A new benchmark for natural language understanding*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4885–4901.

[186] Q. Ning, B. Zhou, Z. Feng, H. Peng, and D. Roth, *CogCompTime: A tool for understanding time in natural language*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, Nov. 2018, Association for Computational Linguistics, pp. 72–77.

[187] T. Niven and H.-Y. Kao, *Probing neural network comprehension of natural language arguments*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 4658–4664.

[188] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, *Universal Dependencies v1: A multilingual treebank collection*, in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, May 2016, European Language Resources Association (ELRA), pp. 1659–1666.

[189] B. Oguz, X. Chen, V. Karpukhin, S. Peshterliev, D. Okhonko, M. Schlichtkrull, S. Gupta, Y. Mehdad, and S. Yih, *Unified Open-Domain Question Answering with Structured and Unstructured Knowledge*, preprint, arXiv, 2020.

[190] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., *Training Language Models to Follow Instructions with Human Feedback*, preprint, arXiv, 2022.

[191] S. Panthaplackel, A. Benton, and M. Dredze, *Updated headline generation: Creating updated summaries for evolving news stories*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 6438–6461.

[192] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer, *An information bottleneck approach for controlling conciseness in rationale extraction*, in Proceedings of the

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 1938–1952.

[193] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, *A decomposable attention model for natural language inference*, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Nov. 2016, Association for Computational Linguistics, pp. 2249–2255.

[194] A. P. Parikh, X. Wang, S. Gehrmann, M. Faruqui, B. Dhingra, D. Yang, and D. Das, *ToTTo: A controlled table-to-text generation dataset*, in Proceedings of EMNLP, Online, Nov. 2020, Association for Computational Linguistics, pp. 1173–1186.

[195] P. Pasupat and P. Liang, *Compositional semantic parsing on semi-structured tables*, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, July 2015, Association for Computational Linguistics, pp. 1470–1480.

[196] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An imperative style, high-performance deep learning library*, in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., Curran Associates, Inc., 2019, pp. 8024–8035.

[197] V. Patil, P. Talukdar, and S. Sarawagi, *Overlap-Based Vocabulary Generation Improves Cross-Lingual Transfer among Related Languages*, preprint, arXiv, 2022.

[198] J. Pennington, R. Socher, and C. Manning, *GloVe: Global vectors for word representation*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Oct. 2014, Association for Computational Linguistics, pp. 1532–1543.

[199] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, *Knowledge enhanced contextual word representations*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 43–54.

[200] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, *Language models as knowledge bases?* in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 2463–2473.

[201] J. Phang, I. Calixto, P. M. Htut, Y. Pruksachatkun, H. Liu, C. Vania, K. Kann, and S. R. Bowman, *English intermediate-task training improves zero-shot cross-lingual transfer too*, in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Suzhou, China, Dec. 2020, Association for Computational Linguistics, pp. 557–575.

[202] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme, *Hypothesis only baselines in natural language inference*, in Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 180–191.

[203] E. M. Ponti, G. Glavaš, O. Majewska, Q. Liu, I. Vulić, and A. Korhonen, *XCOPA: A multilingual dataset for causal commonsense reasoning*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 2362–2376.

[204] A. Pramanick and I. Bhattacharya, *Joint learning of representations for web-tables, entities and types using graph convolutional network*, in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, Apr. 2021, Association for Computational Linguistics, pp. 1197–1206.

[205] Y. Pruksachatkun, J. Phang, H. Liu, P. M. Htut, X. Zhang, R. Y. Pang, C. Vania, K. Kann, and S. R. Bowman, *Intermediate-task transfer learning with pretrained language models: When and why does it work?* in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 5231–5247.

[206] G. Qin and J. Eisner, *Learning how to ask: Querying LMs with mixtures of soft prompts*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 5203–5212.

[207] D. Radev, R. Zhang, A. Rau, A. Sivaprasad, C. Hsieh, N. F. Rajani, X. Tang, A. Vyas, N. Verma, P. Krishna, et al., *DART: Open-Domain Structured Data Record to Text Generation*, preprint, arXiv, 2020.

[208] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving language understanding by generative pre-training*, URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf, 2018.

[209] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, OpenAI Blog, 1 (2019), p. 9.

[210] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al., *Exploring the limits of transfer learning with a unified text-to-text transformer.*, J. Mach. Learn. Res., 21 (2020), pp. 1–67.

[211] A. Rahimi, Y. Li, and T. Cohn, *Massively multilingual transfer for NER*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 151–164.

[212] D. Rajagopal, S. Shakeri, C. N. dos Santos, E. Hovy, and C.-C. Chang, *Counterfactual Data Augmentation Improves Factuality of abstractive Summarization*, preprint, arXiv, 2022.

[213] P. Rajpurkar, R. Jia, and P. Liang, *Know what you don't know: Unanswerable questions for SQuAD*, in Proceedings of the 56th Annual Meeting of the Association for

Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, July 2018, Association for Computational Linguistics, pp. 784–789.

[214] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, *SQuAD: 100,000+ questions for machine comprehension of text*, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Nov. 2016, Association for Computational Linguistics, pp. 2383–2392.

[215] A. Ravichander, Y. Belinkov, and E. Hovy, *Probing the probing paradigm: Does probing accuracy entail task relevance?* in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, Apr. 2021, Association for Computational Linguistics, pp. 3363–3377.

[216] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, *COMET: A neural framework for MT evaluation*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 2685–2702.

[217] N. Reimers and I. Gurevych, *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 3982–3992.

[218] M. T. Ribeir, S. Singh, and C. Guestrin, *Semantically equivalent adversarial rules for debugging NLP models*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 2018, Association for Computational Linguistics, pp. 856–865.

[219] M. T. Ribeiro, S. Singh, and C. Guestrin, *"Why should I trust you?" Explaining the predictions of any classifier*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, California, USA, June 2016, Association for Computing Machinery, pp. 1135–1144.

[220] M. T. Ribeiro, S. Singh, and C. Guestrin, *Anchors: High-precision model-agnostic explanations*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, Apr. 2018, Association for the Advancement of Artificial Intelligence.

[221] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, *Beyond accuracy: Behavioral testing of NLP models with CheckList*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4902–4912.

[222] K. Richardson, H. Hu, L. Moss, and A. Sabharwal, *Probing natural language inference models through semantic fragments*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, Apr. 2020, pp. 8713–8721.

[223] K. Richardson and A. Sabharwal, *What does my QA model know? Devising controlled probes using expert knowledge*, Trans. Assoc. Comput. Linguist., 8 (2020), pp. 572–588.

[224] A. Roberts, C. Raffel, and N. Shazeer, *How much knowledge can you pack into the parameters of a language model?* in Proceedings of the 2020 Conference on Empirical

Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 5418–5426.

[225] O. Rozen, V. Shwartz, R. Aharoni, and I. Dagan, *Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets*, in Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 196–205.

[226] N. X. Ru Wang, D. Mahajan, M. Danilevsky, and S. Rosenthal, *SemEval2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS)*, Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, Association for Computational Linguistics, pp. 317–326.

[227] X. Ruan, M. Jin, J. Ma, H. Yang, L. Jiang, Y. Mo, and M. Zhou, *Sattiy at SemEval-2021 task 9: An ensemble solution for statement verification and evidence finding with tables*, in Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, Aug. 2021, Association for Computational Linguistics, pp. 1255–1261.

[228] S. Ruder, N. Constant, J. Botha, A. Siddhant, O. Firat, J. Fu, P. Liu, J. Hu, D. Garrette, G. Neubig, and M. Johnson, *XTREME-R: Towards more challenging and nuanced multilingual evaluation*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 10215–10245.

[229] M. Saeed, G. Alfarano, K. Nguyen, D. Pham, R. Troncy, and P. Papotti, *Neural re-rankers for evidence retrieval in the FEVEROUS task*, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 108–112.

[230] A. Saxena, S. Chakrabarti, and P. Talukdar, *Question answering over temporal knowledge graphs*, in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, Aug. 2021, Association for Computational Linguistics, pp. 6663–6676.

[231] T. Schick and H. Schütze, *Exploiting Cloze-Questions for few-shot text classification and natural language inference*, in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, Apr. 2021, Association for Computational Linguistics, pp. 255–269.

[232] T. Schïck and H. Schütze, *It's not just size that matters: Small language models are also few-shot learners*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 2339–2352.

[233] T. Sellam, D. Das, and A. Parikh, *BLEURT: Learning robust metrics for text generation*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 7881–7892.

[234] S. Serrano and N. A. Smith, *Is attention interpretable?* in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 2931–2951.

[235] D. Shah, T. Schuster, and R. Barzilay, *Automatic fact-guided sentence modification*, Proceedings of the AAAI Conference on Artificial Intelligence, 34 (2020), pp. 8791–8798.

[236] A. Shankarampeta, V. Gupta, and S. Zhang, *Enhancing tabular reasoning with pattern exploiting training*, in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online only, Nov. 2022, Association for Computational Linguistics, pp. 706–726.

[237] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, *AutoPrompt: Eliciting knowledge from language models with automatically generated prompts*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 4222–4235.

[238] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, *SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training*, preprint, arXiv, 2021.

[239] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, *MPNet: Masked and permuted pre-training for language understanding*, in Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Vancouver, BC, Canada, 2020, Curran Associates Inc.

[240] A. Spangher, X. Ren, J. May, and N. Peng, *NewsEdits: A news article revision dataset and a novel document-level reasoning challenge*, in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, July 2022, Association for Computational Linguistics, pp. 127–157.

[241] L. Specia, F. Blain, M. Fomicheva, E. Fonseca, V. Chaudhary, F. Guzmán, and A. F. T. Martins, *Findings of the WMT 2020 shared task on quality estimation*, in Proceedings of the Fifth Conference on Machine Translation, Online, Nov. 2020, Association for Computational Linguistics, pp. 743–764.

[242] L. Specia, F. Blain, V. Logacheva, R. F. Astudillo, and A. F. T. Martins, *Findings of the WMT 2018 shared task on quality estimation*, in Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, Oct. 2018, Association for Computational Linguistics, pp. 689–709.

[243] R. Speer, J. Chin, and C. Havasi, *Conceptnet 5.5: An open multilingual graph of general knowledge*, in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, San Francisco, California, USA, 2017, AAAI Press, p. 4444–4451.

[244] A. Srinivasan, S. Sitaram, T. Ganu, S. Dandapat, K. Bali, and M. Choudhury, *Predicting the Performance of Multilingual NLP Models*, preprint, arXiv, 2021.

[245] H. Sun, H. Ma, X. He, W.-t. Yih, Y. Su, and X. Yan, *Table cell search for question answering*, in Proceedings of the 25th International Conference on World Wide Web, ACM - Association for Computing Machinery, April 2016, pp. 771–782.

[246] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, *ERNIE: Enhanced Representation through Knowledge Integration*, preprint, arXiv, 2019.

[247] A. Talmor, J. Herzig, N. Lourie, and J. Berant, *CommonsenseQA: A question answering challenge targeting commonsense knowledge*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 4149–4158.

[248] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi, and J. Berant, *MultiModalQA: Complex question answering over text, tables and images*, in International Conference on Learning Representations, ICLR, 2021.

[249] D. Tam, R. R. Menon, M. Bansal, S. Srivastava, and C. Raffel, *Improving and simplifying pattern exploiting training*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 4980–4991.

[250] R. Tanaka, K. Nishida, K. Nishida, T. Hasegawa, I. Saito, and K. Saito, *SlideVQA: A dataset for document visual question answering on multiple images*, in 37th AAAI Conference on Artificial Intelligence, Washington DC, 2023, AAAI Press.

[251] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, *Multilingual translation from denoising pre-training*, in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, Aug. 2021, Association for Computational Linguistics, pp. 3450–3466.

[252] Z. Tang, Z. Yang, G. Wang, Y. Fang, Y. Liu, C. Zhu, M. Zeng, C. Zhang, and M. Bansal, *Unifying Vision, Text, and Layout for Universal Document Processing*, preprint, arXiv, 2022.

[253] I. Tarunes, S. Aditya, and M. Choudury, *Trusting roBERTa over BERT: Insights from Checklisting the Natural Language Inference Task*, preprint, arXiv, 2021.

[254] I. Tarunesh, S. Aditya, and M. Choudhury, *LoNLI: An Extensible Framework for Testing Diverse Logical Reasoning Capabilities for NLI*, preprint, arXiv, 2021.

[255] O. Temiz, Ö. O. Kılıç, A. O. Kızıldağ, and T. Taşkaya Temizel, *A fact checking and verification system for FEVEROUS using a zero-shot learning approach*, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 113–120.

[256] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, *Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 296–310.

[257] M. Trabelsi, Z. Chen, S. Zhang, B. D. Davison, and J. Heflin, *StruBERT: Structure-aware BERT for table search and matching*, in Proceedings of the ACM Web Conference 2022, WWW '22, Virtual Event, Lyon, France, 2022, Association for Computing Machinery, p. 442–451.

[258] C. Tran, S. Bhosale, J. Cross, P. Koehn, S. Edunov, and A. Fan, *Facebook AI's WMT21 news translation task submission*, in Proceedings of the Sixth Conference on Machine Translation, Online, Nov. 2021, Association for Computational Linguistics, pp. 205–215.

[259] L. Tu, G. Lalwani, S. Gella, and H. He, *An empirical study on robustness to spurious correlations using pre-trained language models*, Trans. Assoc. Comput. Linguist., 8 (2020), pp. 621–633.

[260] M. Umair and F. Ferraro, *Transferring Semantic Knowledge into Language Encoders*, preprint, arXiv, 2021.

[261] H. Varma, A. Jain, P. Ratadiya, and A. Rathi, *AttesTable at SemEval-2021 task 9: Extending statement verification with tables for unknown class, and semantic evidence finding*, in Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, Aug. 2021, Association for Computational Linguistics, pp. 1276–1282.

[262] Y. Varun, A. Sharma, and V. Gupta, *Trans-KBLSTM: An external knowledge enhanced transformer BiLSTM model for tabular reasoning*, in Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Dublin, Ireland and Online, May 2022, Association for Computational Linguistics, pp. 62–78.

[263] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, Adv. Neural Inf. Process. Syst., 30 (2017), pp. 5998–6008.

[264] P. Venetis, A. Halevy, J. Madhavan, M. Paşca, W. Shen, F. Wu, G. Miao, and C. Wu, *Recovering semantics of tables on the web*, Proc. VLDB Endow., 4 (2011), p. 528–538.

[265] E. Voita and I. Titov, *Information-theoretic probing with minimum description length*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 183–196.

[266] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, *Universal adversarial triggers for attacking and analyzing NLP*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 2153–2162.

[267] E. Wallace, Y. Wang, S. Li, S. Singh, and M. Gardner, *Do NLP models know numbers? Probing numeracy in embeddings*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 5307–5315.

[268] A. Wang, J. Hula, P. Xia, R. Pappagari, R. T. McCoy, R. Patel, N. Kim, I. Tenney, Y. Huang, K. Yu, S. Jin, B. Chen, B. Van Durme, E. Grave, E. Pavlick, and S. R. Bowman, *Can you tell me how to get past Sesame Street? Sentence-level pretraining beyond language modeling*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 4465–4476.

[269] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, *SuperGLUE: A stickier benchmark for general-purpose language understanding systems*, in Advances in Neural Information Processing Systems, H. Wallach,

H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019, pp. 3266–3280.

[270] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, *GLUE: A multi-task benchmark and analysis platform for natural language understanding*, in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, Nov. 2018, Association for Computational Linguistics, pp. 353–355.

[271] J. Wang, H. Wang, Z. Wang, and K. Q. Zhu, *Understanding tables on the web*, in Proceedings of the 31st International Conference on Conceptual Modeling, ER'12, Florence, Italy, 2012, Springer-Verlag, p. 141–155.

[272] S. Wang, W. Zhong, D. Tang, Z. Wei, Z. Fan, D. Jiang, M. Zhou, and N. Duan, *Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text*, preprint, arXiv, 2021.

[273] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, *KEPLER: A unified model for knowledge embedding and pre-trained language representation*, Trans. Assoc. Comput. Linguist., 9 (2021), pp. 176–194.

[274] X. Wang, P. Kapanipathi, R. Musa, M. Yu, K. Talamadupula, I. Abdelaziz, M. Chang, A. Fokoue, B. Makni, N. Mattei, and M. Witbrock, *Improving natural language inference using external knowledge in the science questions domain*, Proc. AAAI Conf. Artif. Intell., 33 (2019), pp. 7208–7215.

[275] Z. Wang and A. Culotta, *Robustness to spurious correlations in text classification via automatically generated counterfactuals*, in Proc. AAAI Conf. Artif. Intell., vol. 35, 2021, pp. 14024–14031.

[276] Z. Wang, Z. Li, J. Li, J. Tang, and J. Z. Pan, *Transfer learning based cross-lingual knowledge extraction for Wikipedia*, in Association for Computational Linguistics, 2013, pp. 641–650.

[277] C. Wei, S. M. Xie, and T. Ma, *Why Do Pretrained Language Models Help in Downstream Tasks? An Analysis of Head and Prompt Tuning*, preprint, CoRR, 2021.

[278] H. Wen, Y. Qu, H. Ji, Q. Ning, J. Han, A. Sil, H. Tong, and D. Roth, *Event time extraction and propagation via graph attention networks*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 62–73.

[279] S. Wiegreffe and Y. Pinter, *Attention is not not explanation*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 11–20.

[280] A. Williams, N. Nangia, and S. Bowman, *A broad-coverage challenge corpus for sentence understanding through inference*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 1112–1122.

[281] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, *Transformers: State-of-the-art natural language processing*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Oct. 2020, Association for Computational Linguistics, pp. 38–45.

[282] X. Wu, J. Zhang, and H. Li, *Text-to-table: A new way of information extraction*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 2518–2533.

[283] J. Xia, C. Wu, and M. Yan, *Incorporating relation knowledge into commonsense reading comprehension with multi-task learning*, in Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2393–2396.

[284] W. Xiong, J. Du, W. Y. Wang, and V. Stoyanov, *Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model*, in 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020, ICLR.

[285] Y. Xu, C. Zhu, R. Xu, Y. Liu, M. Zeng, and X. Huang, *Fusing context into knowledge graph for commonsense question answering*, in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, Aug. 2021, Association for Computational Linguistics, pp. 1201–1207.

[286] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, *mT5: A massively multilingual pre-trained text-to-text transformer*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 483–498.

[287] V. Yadav, S. Bethard, and M. Surdeanu, *Alignment over heterogeneous embeddings for question answering*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 2681–2691.

[288] V. Yadav, S. Bëthard, and M. Surdeanu, *Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4514–4525.

[289] J. Yang, A. Gupta, S. Upadhyay, L. He, R. Goel, and S. Paul, *TableFormer: Robust transformer modeling for table-text encoding*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, Association for Computational Linguistics, pp. 528–537.

[290] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge, *PAWS-X: A cross-lingual adversarial dataset for paraphrase identification*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 3687–3692.

[291] P. Yin, G. Neubig, W.-t. Yih, and S. Riedel, *TaBERT: Pretraining for joint understanding of textual and tabular data*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 8413–8426.

[292] W. Yin, D. Radev, and C. Xiong, *DocNLI: A Large-Scale Dataset for Document-Level Natural Language Inference*, preprint, arXiv, 2021.

[293] O. Yoran, A. Talmor, and J. Berant, *Turning Tables: Generating Examples from Semi-Structured Tables for Endowing Language Models with Reasoning Skills*, preprint, arXiv, 2021.

[294] T. Yu, C.-S. Wu, X. V. Lin, B. Wang, Y. C. Tan, X. Yang, D. Radev, R. Socher, and C. Xiong, *GraPPa: Grammar-augmented pre-training for table semantic parsing*, in International Conference of Learning Representation, 2021, ICLR.

[295] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, et al., *Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Oct.–Nov. 2018, Association for Computational Linguistics, pp. 3911–3921.

[296] W. Zaremba, I. Sutskever, and O. Vinyals, *Recurrent Neural Network Regularization*, preprint, arXiv, 2014.

[297] V. Zayats, K. Toutanova, and M. Ostendorf, *Representations for Question Answering from Documents with Tables and Text*, preprint, arXiv, 2021.

[298] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, *SWAG: A large-scale adversarial dataset for grounded commonsense inference*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Oct.–Nov. 2018, Association for Computational Linguistics, pp. 93–104.

[299] C. Zhang, J. Zhao, H. Zhang, K.-W. Chang, and C.-J. Hsieh, *Double perturbation: On the robustness of robustness and counterfactual bias evaluation*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 3899–3916.

[300] H. Zhang, Y. Wang, S. Wang, X. Cao, F. Zhang, and Z. Wang, *Table fact verification with structure-aware transformer*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 1624–1629.

[301] L. Zhang, S. Zhang, and K. Balog, *Table2Vec: Neural word and entity embeddings for table population and retrieval*, in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Paris, France, 2019, Association for Computing Machinery, p. 1029–1032.

[302] S. Zhang and K. Balog, *Web table extraction, retrieval, and augmentation: A survey*, ACM Trans. Intell. Syst. Technol., 11 (2020), pp. 13:1–13:35.

[303] S. Zhang, Z. Dai, K. Balog, and J. Callan, *Summarizing and exploring tabular data in conversational search*, in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Virtual Event, China, 2020, Association for Computing Machinery, pp. 1537–1540.

[304] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *BERTScore: Evaluating text generation with BERT*, in International Conference on Learning Representations, 2019.

[305] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, *ERNIE: Enhanced language representation with informative entities*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 1441–1451.

[306] M. Zhao, F. Mi, Y. Wang, M. Li, X. Jiang, Q. Liu, and H. Schuetze, *LMTurk: Few-shot learners as crowdsourcing workers in a language-model-as-a-service framework*, in Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, United States, July 2022, Association for Computational Linguistics, pp. 675–692.

[307] Z. Zhao, D. Dua, and S. Singh, *Generating natural adversarial examples*, in International Conference on Learning Representations, Brussels, Belgium, Oct.–Nov. 2018, Association for Computational Linguistics, pp. 2890–2896.

[308] B. Zhou, D. Khashabi, Q. Ning, and D. Roth, *"going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 3363–3369.

[309] B. Zhou, K. Richardson, Q. Ning, T. Khot, A. Sabharwal, and D. Roth, *Temporal reasoning on implicit events from distant supervision*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 1361–1371.

[310] X. Zhou and M. Bansal, *Towards robustifying NLI models against lexical dataset biases*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 8759–8771.