Inference and Reasoning on Semi-Structured Tables

Vivek Gupta
Kahlert School of Computing, University of Utah

1 Introduction

The goal of my research is to design, implement and analyze models for *semi-structured tabular data* (cf. Figure 1). Although neural network models has gained success on unstructured text (sentences and paragraph), their reasoning capacity on semi-structured text is poorly understood. Consequently, people (even NLP experts) have little perception on how NLP models reasons. Semi-structured data can be utilized to increase understanding about model reasoning ability on textual information. While working with tabular data, the following questions were addressed:

Q1. How do models designed for unstructured text adapt to tabular data?(§2.1) The infobox table (c.f. Figure 1) does not explicitly state the relationship between the keys and values (Gupta et al., 2020b). Additionally, only a portion of the table is necessary for the model to predict; the remainder acts as a information pollution (Neeraja et al., 2021; Gupta et al., 2022b). All keys except *"Length"* and *"Producer"* are unrelated to hypothesis H1. Irrelevant rows could produce erroneous correlations, leading to correct predictions for the wrong reasons.

Q2. How do one incorporate knowledge into tabular reasoning models?(§2.2) Tables often 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, one need to interpret the key *'Length'* n 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 (Neeraja et al., 2021; Varun et al., 2022). This affects interpreting meaning of words such as "*less than*" in H1 (c.f. Figure 1).

Q3. How to ensure that the model is doing correct evidence-based reasoning?(§2.3) Recent studies show that deep learning systems are brittle and memorize spurious patterns such as annotation artefacts, often amplify societal biases (Bolukbasi et al.; Zhao et al., 2017; Poliak et al., 2018; Niven and Kao, 2019). As a result, the model suffers from a lack of information trustworthiness. We investigate this issue in the context of tables, and developed several systematic logical probes (Gupta et al., 2022a) to evaluate models' reasoning abilities in terms of correct evidence selection (Gupta et al., 2022b), robustness to input perturbation particularly hypothesis, and reasoning ability over counterfactual information.

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: A semi-structured premise (the table 'Breakfast in America') example from (Gupta et al., 2020b). Hypotheses H1 are entailed by it, H2 is neither entailed nor contradictory, and H3 is a contradiction. The 'Relevance' column shows the hypotheses that use the corresponding row for reasoning. The colored text (and superscripts) in the table and hypothesis highlights relevant rows.

Q4. How do models reason about dynamic information, particularly temporal information? Numerous data pieces inside an entity evolve and change throughout time (c.f. §3.1). For instance, a country's population, water resources or it's official representatives change frequently. Robust models must consider this changes and ensure that they can reason well across temporal dimensions.

2 Current Work

2.1 How do models designed for unstructured text adapt to tabular data?

To study this questions we created InfoTabS, a semi-structure tabular inference dataset. InfoTabS[1], comprising of human-written textual hypotheses based on premises that are extracted from Wikipedia info-boxes. InfoTabS consists of $23,738$ premise-hypothesis pairs, whose premises are based on Wikipedia infoboxes. Figure 1 shows an example table from the dataset with three hypotheses. The dataset contains $2,540$ distinct infoboxes representing a variety of domains. All hypotheses were written and labeled by MTurk workers. InfoTabS incorporates several diverse kinds of reasoning (numerical, temporal, knowledge and common sense etc.) all adapted from the Glue (Wang et al., 2018) and SuperGlue (Wang et al., 2019) benchmarks, which are typically missing in earlier NLI datasets. For example, in Figure 1, consider the hypothesis sentence H1. In order to determine whether the hypothesis entails the premise, one needs to look up multiple rows (*'Length'* and *'Producer'*), conclude that *'Length'* in Album terms denotes the total length of the album's songs (i.e. Album Singles), and *'46:06'* where the album length is in minutes rather than an hour (using common sense). In addition to the regular training and development sets, to differentiate models' true learning ability from learning spurious correlated patterns in the data (artifacts), we created three challenge test sets of equal size. The α_1 set (200 tables, 1800 table-hypothesis pair) represents a standard test set that is 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.

[1] InfoTabS website: `https://infotabs.github.io`

We use a universal template “The *row-key* of *table-title* are *row-values*." to represent the table as a paragraph i.e. *Table as a Paragraph*, with each row representing a sentence. The penultimate row of Figure 2 table presents the performance of the model trained on the entire training data, while the last row presents the performance of the 5xCV models. The table also shows the hypothesis-only baseline (Poliak et al., 2018; Gururangan et al., 2018) and human agreement on the labels[2]. We found that existing state-of-the-art models, e.g., RoBERTa-Large for NLI, underperform on INFOTABS dataset compared to the majority human agreement performance, suggesting that reasoning about tables can pose a difficult modeling challenge. Recently, we also extend the INFOTABS to it’s multilingual version XINFOTABS (Minhas et al., 2022; Agarwal et al., 2022), 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.

Model	α_1	α_2	α_3
Human	**84.04**	**83.88**	**79.33**
Hypothesis Only	60.48	48.26	48.89
RoBERTa$_{\text{LARGE}}$	74.88	65.55	64.94
5xCV	72.41$_{(1.4)}$	63.02$_{(1.9)}$	61.82$_{(1.4)}$

Figure 2: Results of the *Table as a Paragraph* strategy on INFOTABS subsets with RoBERTa$_L$ model, hypothesis-only baseline and majority human agreement. The last row represents the average performances (and s.t.d as subscripts) using models obtained via cross validation.

2.2 How do we incorporate knowledge into tabular reasoning models?

We examined this question via the lens of effective pre-processing techniques[3], as described below: **Better table representation**: The table does not explicitly state the relationship between the keys and values. §2.1 suggested the use of a universal template which leads to most sentences being ungrammatical, e.g., "*The recorded of Breakfast in America is 29 March 1998.*". To address this, we propose using entity specific templates **(BTR)** by using value entity types DATE or MONEY or CARDINAL or BOOL. The final sentence now become grammatically correct, e.g., "*Breakfast in America was recorded on March 29th, 1998.*". Furthermore, we also add category-specific information, e.g., "*Breakfast in America is an album.*".

Missing implicit lexical knowledge: This affects interpreting meaning of words like *‘less than’*, and *‘most of’* in H1 and H2 respectively. Limited training data affects the interpretation of *synonyms*, *antonyns*, *hypernyms*, *Hyponyns*, and *Co-hyponyms* words such as “fewer", “over", “more than", “less than", “over", “under", “negations", and others. We find that pre-training on a large Natural Language Inference dataset helps expose the model to diverse lexical constructions and make representation tuned to the NLI task. So firstly, we intermediately pre-train with MNLI data **(KG implicit)** and then subsequently fine tune on the tabular inference INFOTABS dataset.

Presence of distracting information: Only select rows are relevant for a given hypothesis. For example, the key *‘Recorded’* is relevant for the hypothesis H2 and H3 but irrelevant for the hypothesis H3. Furthermore, due to BERT tokenization limit, useful rows in longer tables might be cropped. To handle this we propose, Distracting Row Removal **(DRR)**, where we select only rows relevant to the hypothesis. For this, we adopt the Alignment based retrieval algorithm with fastText vectors as detailed in Yadav et al. (2019, 2020). For example, we prune the table with only rows *‘Length’* and *‘Producer’* for hypothesis H1. We also explore the sensitivity to extraction method and introduce *trustworthy tabular inference* (Gupta et al., 2022b). In, *trustworthy tabular inference*, we split the NLI task into causal sequential task of evidence extraction and inference on extracted evidence. We utilize several supervised and unsupervised methods for the evidence extraction.

Premise	α_1	α_2	α_3
Human	**84.04**	**83.88**	**79.33**
Paragraph	74.88	65.55	64.94
BTR	75.29	66.50	64.26
+KG implicit	78.27	71.87	66.77
+DRR	78.13	70.90	68.98
+KG explicit	**78.42**	**71.97**	**70.03**

Figure 3: Accuracy with the proposed modifications on the test sets. Here, + represents the change with respect to the previous row.

Missing domain knowledge about keys: We need to interpret the meaning of the table key in the correct context for this table. In the case of H1, we need to interpret *’Length’* in the album context. For example, here, the length must be interpreted as in *album* context: "*The such of total playtime of all the songs in the record album.*" rather as "*The dimension of the larger side of a portrait.*" as in *painting* context. We append explicit information **(KG explicit)** to enrich the keys. Explicit knowledge helps improve the model’s ability to disambiguate the meaning of the keys. We utilize BERT on wordnet examples to get key embeddings, then use BERT to get key embeddings from the premise, and finally select the best match, based on similarity score, and add its definition to the premise.

Our proposed effective pre-processing approaches lead to substantial improvements in prediction quality, especially on adversarial α_2 and α_3 test sets as shown in Figure 3 Table[4]. Definitions are lengthier and sometime unnecessary bring information pollution. To address this, we recently propose structured knowledge usage from factual and commonsense knowledge graphs such as DBpedia, ATOMIC and ConceptNet. We introduce efficient in-modeling knowledge incorporation via combining Bi-LSTM with transformer, i.e. TransKBLSTM (Varun et al., 2022). The proposed solutions are also applicable to question answering and generation problems with both tabular and textual inputs.

[2] RoBERTa$_L$ outperformed other pre-trained embeddings in development set testing. BERT$_B$, RoBERTa$_B$, BERT$_L$, ALBERT$_B$, and ALBERT$_L$ had development set accuracies of 63.0%, 67.23%, 69.34%, 70.44%, and 70.88%. Given the substantial computational costs, we have not duplicated our studies on these additional models, but we anticipate the findings to generalize. [3] Knowledge-INFOTABS website: `https://knowledge-infotabs.github.io/` [4] Three random seed runs were averaged with a standard deviation of 0.33 (+KG explicit), 0.46 (+DRR), 0.61 (+KG implicit), 0.86 (BPR), over all sets. All improvements are statistically significant with $p < 0.05$, except α_1 for BPR representation w.r.t to Para (Original).

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

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[5]. Overall, the guiding premise for this (in-/co-)variants perturbation work 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, described below using example in Figure 1.

(a.) 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 H1 of Figure 1 if the token "less than" is replaced with "more than", the model prediction should change from ENTAIL to CONTRADICT. To create such probe, we identify a set of reasoning categories and characterize the relationship between a tabular premise and a hypothesis.

From the analysis of artifact probe, we found that the model heavily relies on correlations between a hypothesis' sentence structure and its label. Thus, 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 non-tabular tasks such as question answering, sentiment analysis, document classification, natural language inference, etc. (e.g. Ribeiro et al., 2020; Richardson et al., 2020; Goel et al., 2021; Lewis et al., 2021; Tarunesh et al., 2021).

(b.) 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 *Length* and *Producer* rows of Figure 1. 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. We define four types of table modifications that are agnostic to the relevance of rows to a hypothesis: (a) **row deletion**, i.e. deleting information, (b) **row insertion**, i.e. inserting new information, (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.[6] Overall from evidence-selection probing, we found the model does not look at correct evidence (Figure 4) for correct reasoning and rather leverages spurious patterns and statistical correlations to make predictions. A recent study by Lewis et al. (2021) on non-tabular question-answering shows that models indeed leverage spurious patterns to answer a large fraction (60-70%) of questions.

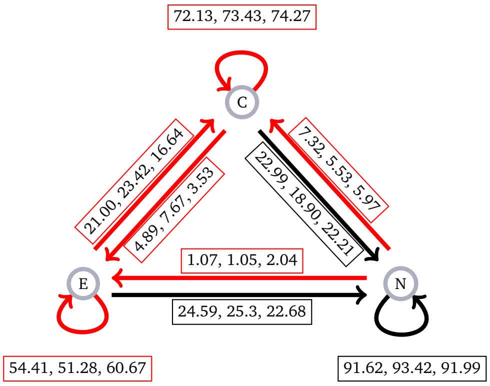

Figure 4: Changes in model predictions after deletion of relevant rows. Directed edges are labeled with transition percentages from the source node label to the target node label. The number triple corresponds to α_1, α_2 and α_3 test sets respectively and for each source node, adds up to 100% over the outgoing edges. Red lines represent invalid transitions while **black** lines represent valid transitions.

(c.) 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 and year of the *Released* date changed to "December" and "1978 "respectively, then the model should change the label of H3 in Figure 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. We used similar techniques for synthetic and counterfactual tabular augmentation data generation (Kumar et al., 2022) to enhance tabular reasoning.

From counterfactual probes, 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 Clark and Etzioni (2016); Jia and

[5] This strategy has been either explicitly or implicitly also employed for recent non-tabular work work (Ribeiro et al., 2020; Gardner et al., 2020).

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

Liang (2017); Kaushik et al. (2020); Huang et al. (2020); Gardner et al. (2020); Tu et al. (2020); Liu et al. (2021); Zhang et al. (2021); Wang et al. (2021) for unstructured text. We refer the reader to the Gupta et al. (2022a) for probes details and more results. Additionally, we also released a interactive annotation platform (Jain et al., 2021) for generating effective tabular perturbations.

3 Ongoing Research Direction

3.1 How do models reason about dynamic information, particularly temporal information?

Numerous components of information about an entity evolve throughout time. E.g., from the Wikipedia Infobox of current POTUS "Joe Biden" [7] one can observe temporal variation in entity information across several relational aspects (keys) such as, e.g., *official position*, *marital status*, *political affiliation*, and others. Using the information present in the InfoBox one can easily reason over several interesting temporal questions as shown in Figure 5[8].

Question Types. To handle such temporal variation and challenging questions, an NLP model should be capable of reasoning across time (i.e., temporal alteration). We propose to construct a novel dataset of table-based question answers revolving around temporal questions. The dataset would contain an Infobox table and multiple questions revolving around time. The questions will be broadly of the following three types: (a) the question contain temporal aspect and answer seek relational keys (e.g. Q1.). (b) the question contains relational keys and the answer seeks time (e.g. Q4.).

Q1. What was Biden's political affiliation in 1953? ; A1. Independent
Q2. What position did Biden hold in year 2012?; A2: VPOTUS
Q3. How many positions did Biden's hold in 2009?; A3: Four
Q4. How long has Biden chaired the Senate Foreign Relations Committee?; A4: Six Years

Figure 5: Temporal Reasoning over InfoBox tables.

Temporal Reasoning Types. We also plan to include varying levels of reasoning difficulty in the questions as follows: (a) mention times in questions that is explicitly mention in the entity tables. (b) mention time in the questions, which can be inferred implicitly from the tables. (c) questions involve reasoning across multiple relational keys across time which vary temporally. (d) mentions relational keys explicitly or implicitly in the question with an answer involving temporal reasoning. (e) mentions temporal relationships such as 'within', 'between', 'before', 'after', etc. and seeks either a relational key or value as the answer. and (f) any other complex temporal reasoning questions. To answer such implicit commensense based temporal questions, one require neural modular and neuro-symbolic modeling (Gupta et al., 2020a; Li et al., 2019), question decomposition's and distinct supervision approaches (Zhou et al., 2021, 2022).

Data Construction. For dataset construction, we first plan to re-purpose/recast existing temporal questions answering datasets (Jena et al., 2022) such as (a) Time-Sensitive-QA (Chen et al., 2021b), TORQUE (Ning et al., 2020) entity-specific reading comprehension dataset with time-sensitive questions over paragraph taken from wikipedia pages, (b) TempQA-WD (Neelam et al., 2022), CRONQUESTIONS (Saxena et al., 2021), TempQuestions (Jia et al., 2018a) question answering datasets over knowledge graph embedding with temporal link. In all cases, we plan to replace the Wikipedia paragraph or knowledge graph with the Wikipedia infobox table. We also plan to explore questions form open domain (Zhang and Choi, 2021) and cloze-form (Dhingra et al., 2022) or event-centric (Ning et al., 2018; Wen et al., 2021; Chen et al., 2021a) temporal questing answering datasets. Additionally, we intend to use the mturk platform to crowdsource difficult template queries that are not addressed by recasting. We will also utilize crowdsourcing for manual and automated paraphrasing (Zhao et al., 2021) to further rephrase these templates for additional lexical variation (Kumar et al., 2022).

Q1. What is the water area of Salt Lake City in year 2020?; A1. 0.47 sq mi (1.22 km2);
Q2. Between 2010 and 2020, how much did SLC's population grow?; A2: 13283
Q3. Who was the mayor of Salt Lake City in year 2010? A3. Jackie Biskupski (D)
Q4. What is the net change is population rank of SLC from 2010 to 2020?; A4: +6
Q5. What is the city area of SLC in year 2020?; A5: 110.81 sq-mi (286.99 km^2)

Figure 6: Dynamic Temporal Reasoning over temporally varied InfoBox tables.

Benchmark Models. The purpose of the proposed dataset is to study any model's temporal reasoning and understanding ability in a grounded context of the form of succinct semi-structured data such as entity tables. We intend to investigate temporally tuned language models trained on knowledge-based question answering datasets like CRONKBQA (Saxena et al., 2021), TEQUILA (Jia et al., 2018b), EXAQT (Jia et al., 2021), OTR-QA(Shang et al., 2021), TempoQR(Mavromatis et al., 2021), and others. We also plan to explore methods that incorporate temporal aspects while language model pre-training (Dhingra et al., 2022; Logan IV et al., 2021), rather than fine-tuning on the downstream task.

Other Directions. We anticipate that our proposed approach will spark future research on other related questions, including (a) Table Retrieval Based Temporal Question Answering: This task is a natural open-domain extension of the proposed work, where correct tables need to be retrieved before applying temporal reasoning (b) Dynamic temporal variation across varied temporal tables: this setting explore an entity-table whose information is gradually varied across time. E.g. using the infobox table of *Salt Lake City* of *year 2010* and *year 2020* to answer the challenging questions as shown in Figure 6[9]. (c) Information updates across multilingual tables: this problem explore the table knowledge gap across multilingual tables, using a translation-based alignment method and a reasoning-based updating algorithm.

[7] https://en.wikipedia.org/wiki/Joe_Biden [8] more detailed examples https://bit.ly/3LHMFrx [9] more detailed examples: https://bit.ly/3qZ5DlA

References

Chaitanya Agarwal, Vivek Gupta, Anoop Kunchukuttan, and Manish Shrivastava. 2022. 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*, pages 4018–4037, Seattle, United States. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*.

Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021a. Event-centric natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021b. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Peter Clark and Oren Etzioni. 2016. My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI. *AI Magazine*, 37(1):5–12.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. 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*, pages 42–55, Online. Association for Computational Linguistics.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020a. Neural module networks for reasoning over text. In *International Conference on Learning Representations*.

Vivek Gupta, Riyaz A. Bhat, Atreya Ghosal, Manish Shrivastava, Maneesh Singh, and Vivek Srikumar. 2022a. Is my model using the right evidence? systematic probes for examining evidence-based tabular reasoning. *Transactions of the Association for Computational Linguistics*, 10.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020b. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji, and Vivek Srikumar. 2022b. Right for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the 2022 Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. 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)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

William Huang, Haokun Liu, and Samuel R. Bowman. 2020. 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*, pages 82–87, Online. Association for Computational Linguistics.

Nupur Jain, Vivek Gupta, Anshul Rai, and Gaurav Kumar. 2021. TabPert : An effective platform for tabular perturbation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 350–360, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aashna Jena, Vivek Gupta, Manish Shrivastava, and Julian Eisenschlos. 2022. Leveraging data recasting to enhance tabular reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4483–4496, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1057–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 1807–1810, New York, NY, USA. ACM.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 792–802.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Dibyakanti Kumar, Vivek Gupta, Soumya Sharma, and Shuo Zhang. 2022. Realistic data augmentation framework for enhancing tabular reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4411–4429, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. 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*, pages 1000–1008, Online. Association for Computational Linguistics.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. 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)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert L Logan IV, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2021. Fruit: Faithfully reflecting updated information in text. *arXiv preprint arXiv:2112.08634*.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N. Ioannidis, Soji Adeshina, Phillip R. Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. 2021. Tempoqr: Temporal question reasoning over knowledge graphs.

Bhavnick Minhas, Anant Shankhdhar, Vivek Gupta, Divyanshu Aggarwal, and Shuo Zhang. 2022. XInfoTabS: Evaluating multilingual tabular natural language inference. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 59–77, Dublin, Ireland. Association for Computational Linguistics.

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, et al. 2022. A benchmark for generalizable and interpretable temporal question answering over knowledge bases. *arXiv preprint arXiv:2201.05793*.

J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. 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*, pages 2799–2809, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8713–8721.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. 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)*, pages 6663–6676, Online. Association for Computational Linguistics.

Chao Shang, Peng Qi, Guangtao Wang, Jing Huang, Youzheng Wu, and Bowen Zhou. 2021. Open temporal relation extraction for question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Trusting RoBERTa over BERT: Insights from checklisting the natural language inference task. *arXiv preprint arXiv:2107.07229. Version 1.*

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Yerram Varun, Aayush Sharma, and Vivek Gupta. 2022. 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*, pages 62–78, Dublin, Ireland and Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. 2021. Logic-driven context extension and data augmentation for logical reasoning of text. *arXiv preprint arXiv:2105.03659. Version 1.*

Haoyang Wen, Yanru Qu, Heng Ji, Qiang Ning, Jiawei Han, Avi Sil, Hanghang Tong, and Dan Roth. 2021. 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*, pages 62–73, Online. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. 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)*, pages 2681–2691, Minneapolis, Minnesota. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4514–4525, Online. Association for Computational Linguistics.

Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. 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*, pages 3899–3916, Online. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Mengjie Zhao, Fei Mi, Yasheng Wang, Minglei Li, Xin Jiang, Qun Liu, and Hinrich Schütze. 2021. Lmturk: Few-shot learners as crowdsourcing workers. *arXiv preprint arXiv:2112.07522*.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. 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*, pages 1361–1371, Online. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. 2022. Learning to decompose: Hypothetical question decomposition based on comparable texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2223–2235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.