Inference and Reasoning on Semi-Structured Tables

Vivek Gupta
Kahlert School of Computing, University of Utah

1 Introduction

Semi-structured tables are a ubiquitous feature in various domains, including e-commerce 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 require complex reasoning and inference to understand the implicit connections across table cells.

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 experts) have limited perception on how models reason. I believe that studying semi-structured data is essential for understanding model reasoning ability on textual information. Therefore, my research focuses on *semi-structured tabular data* (cf. Figure 1). While working with tables, I addressed the following questions:

Q1. How do models designed for unstructured text adapt to (semi-)structured data?(§2.1) Unstructured text explicitly mentions connection between the entities in the sentence/paragraph. However, in (semi-)structured text (e.g. tables) these relationships are latent due to its underlying implicit structure. Furthermore, tables hold information in succinct form, which makes information navigation in the cluttered world challenging (Neeraja et al., 2021; Gupta et al., 2022b).

Q2. How does one incorporate knowledge into tabular models?(§2.2) AI programs that are trained on tables might not understand certain words and phrases, which can make it hard to interpret the tabular information correctly. For example, in a table that lists music albums, the label "*Length*" might not make sense without more information about the context of the table.

Q3. How to ensure that the model is doing correct evidence-based reasoning?(§2.3) AI models suffer from a lack of output trustworthiness, making it difficult to be deployed in the real world. Recent studies show that AI 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). I studied these biases for tables using logical probes (Gupta et al., 2022a).

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 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.

2 My research

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

To study this questions we created INFOTABS (Gupta et al., 2020), a semi-structure tabular inference dataset. INFOTABS consists of human-written textual hypotheses based on premises extracted from Wikipedia info-boxes. Figure 1 shows an example from the INFOTABS dataset, a table with three hypotheses. The dataset contains $2,540$ distinct infoboxes (≈ 24K pairs) representing a variety of domains. INFOTABS incorporates several diverse kinds of reasoning (numerical, temporal, knowledge and common sense etc.) most 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. 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 from topics not present in the training set.

We also created the first set of baselines on INFOTABS dataset. The third row (Universal Encoding) of Figure 2 table presents the performance of the model trained on training data. The table also shows the hypothesis-only baseline (Poliak et al., 2018; Gururangan et al., 2018) and human agreement on the labels. We found that existing inference models, e.g., RoBERTa-LARGE, underperform on INFOTABS compared to the majority human agreement performance, suggesting that reasoning about tables can pose a difficult modeling challenge.

Model	α_1	α_2	α_3
Human	**84.04**	**83.88**	**79.33**
Hypothesis Only	60.48	48.26	48.89
Universal Encoding	74.88	65.55	64.94
Type-Base Encoding	75.29	66.50	64.26
+++Knowledge	**78.42**	**71.97**	**70.03**

Figure 2: Results on INFOTABS representation with RoBERTa$_L$ model, hypothesis-only baseline and majority human agreement. Table also show accuracy with the proposed modifications (§2.2).

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.

2.2 How do we incorporate knowledge into tabular reasoning models?

Tables often lack the necessary context to explain the relationship between different elements, like table attributes and values. As a result, models trained on tables often have difficulty with correct reasoning. To overcome this issue, one approach is to incorporate knowledge through pre-processing (Neeraja et al., 2021).

(a.) **Type-based representation.** A model should understand implicit relationship between table entries. The table does not explicitly state the relationship between the attributes and values. §2.1 suggested the use of a universal template to address this, but this leads to most sentences being incoherent and ungrammatical, e.g., "*The recorded of Breakfast in America is 29 March 1998.*". Incoherent sentences can often limit a model's ability to understand information. To address this, we propose using entity-type specific templates 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.*".

(b.) **Adding lexical knowledge.** Model's should be able to decipher the diverse lexical constructions. A accurate model can distinguish differences between word meanings, such as *"less than"* in H1 and *"most of"* in H2. However, limited training data often affects the model 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 model representation tuned to the NLI task. So firstly, we intermediately pre-train with MNLI data **(implicit knowledge)** and then subsequently fine tune on the tabular inference INFOTABS dataset.

(c.) **Removing distracting information.** A good model should be able to select the pertinent evidence for accurate reasoning. 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. Models can struggle with selecting the right evidence due to the vast amount of surrounding information. To handle this we propose, **distracting row removal**, 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). 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.

(d.) **Adding domain knowledge.** The model needs to understand what the table attribute means in respect to table domain. For example, in H1, the *"Length"* attribute should be understood as *"the total playtime of a music album"*, not as *"the size of the larger side of a portrait"*, which would be the meaning in a *"painting"* domain. To help the model, we provide extra information **(explicit knowledge)** that explains the correct meaning of the attribute. This extra knowledge helps the model to choose the right meaning of the table attribute. We use BERT (Devlin et al., 2019) attribute embeddings to compare wordnet examples with the table premise and add the correct definition as extra context to the premise.

Our proposed knowledge addition approach (+++ knowledge) lead to substantial improvements in prediction quality, especially on adversarial α_2 and α_3 test sets as shown in Figure 2 Table. Definitions can be long and sometimes add unnecessary information, causing confusion. To solve this problem, we suggest using structured knowledge from factual and commonsense knowledge graphs like DBpedia (Auer et al., 2007), ATOMIC (Sap et al., 2019), and ConceptNet (liu, 2004). Our proposed solution, TransKBLSTM (Varun et al., 2022), combines Bi-LSTM with transformer to efficiently incorporate knowledge within the model. This approach can also be used for question answering and generation tasks that involve both tabular and textual inputs.

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[1]. 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).

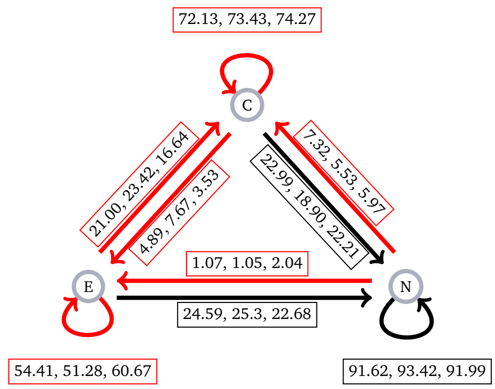

Figure 3: Changes in model predictions after deletion of relevant rows. Directed edges are labeled with transition percentages from the source node label to the target node label. The number triple corresponds to α_1, α_2 and α_3 test sets respectively and for each source node, adds up to 100% over the outgoing edges. Red lines represent invalid transitions while **black** lines represent valid transitions.

(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.[2] Overall from

[1] This strategy has been either explicitly or implicitly also employed for recent non-tabular work (Ribeiro et al., 2020; Gardner et al., 2020). [2] In performing these modifications, we ensure that the modified table does not become inconsistent or self-contradicting.

evidence-selection probing, we found the model does not look at correct evidence (Figure 3) 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.

(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 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 Next Steps

For future, I envision to explore reasoning over (a.) dynamic, (b.) multilingual, and (c.) multi-modal information, in context of semi-structured data. In particular, I wish to explore the following questions:

(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 (Zhou et al., 2021; Neelam et al., 2022; Saxena et al., 2021; Jia et al., 2018; Dhingra et al., 2022; Ning et al., 2018; Wen et al., 2021; Chen et al., 2021, and soon).

(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, I propose utilizing 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 (Iv et al., 2022), news editing (Spangher et al., 2022), headline updation (Panthaplackel et al., 2022), and sentence updation (Shah et al., 2020; Dwivedi-Yu et al., 2022).

(c.) **Navigating Multi-modal Information.** My current work involves studying unimodal tables with simple text. However, I'm keen to expand my research to include multimodal tables with text, symbols, images, and complex nested structures. ***How can model reason on complex multimodal tables?*** I aim to 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 (Liu et al., 2022b; Lee et al., 2022; Liu et al., 2022a), QA on infographicVQA (Mathew et al., 2022; Tanaka et al., 2023), and image-table-text generation (Gatti et al., 2022; Talmor et al.).

By tackling the broader problems of dynamic, multilingual, and multi-modal information in semi-structured data, I hope to contribute to the development of novel methods for reasoning with changing information, and ultimately advance our understanding of these complex data types.

References

2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

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.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

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. 2021. Event-centric natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, and Fabio Petroni. 2022. Editeval: An instruction-based benchmark for text improvements. *arXiv preprint arXiv:2209.13331*.

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.

Prajwal Gatti, Anand Mishra, Manish Gupta, and Mithun Das Gupta. 2022. VisToT: Vision-augmented table-to-text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9936–9949, Abu Dhabi, United Arab Emirates. 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.

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. 2020. 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.

Robert Iv, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2022. 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*, pages 3670–3686, Seattle, United States. 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.

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. 2018. 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.

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.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2022. Pix2struct: Screenshot parsing as pretraining for visual language understanding. *arXiv preprint arXiv:2210.03347*.

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.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2022a. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2022b. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. *arXiv preprint arXiv:2212.09662*.

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.

Minesh Mathew, Viraj Bagal, Ruben Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591. IEEE Computer Society.

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, 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.

Sheena Panthaplackel, Adrian Benton, and Mark Dredze. 2022. 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)*, pages 6438–6461, Dublin, Ireland. 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.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. 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*, pages 3027–3035.

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.

Darsh Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8791–8798.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. 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*, pages 127–157, Seattle, United States. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. *arXiv preprint arXiv:2301.04883*.

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.

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.

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.

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.