Natural Language Engineering (2023), **29**, pp. 1–31
doi:10.1017/S1351324922000171

CAMBRIDGE UNIVERSITY PRESS

SURVEY PAPER

A survey of methods for revealing and overcoming weaknesses of data-driven Natural Language Understanding

Viktor Schlegel*, Goran Nenadic and Riza Batista-Navarro

Department of Computer Science, University of Manchester, Manchester M13 9PL, UK
*Corresponding author. E-mail: viktor.schlegel@manchester.ac.uk

(Received 6 May 2021; revised 10 March 2022; accepted 14 March 2022; first published online 22 April 2022)

Abstract

Recent years have seen a growing number of publications that analyse Natural Language Understanding (NLU) datasets for superficial cues, whether they undermine the complexity of the tasks underlying those datasets and how they impact those models that are optimised and evaluated on this data. This structured survey provides an overview of the evolving research area by categorising reported weaknesses in models and datasets and the methods proposed to reveal and alleviate those weaknesses for the English language. We summarise and discuss the findings and conclude with a set of recommendations for possible future research directions. We hope that it will be a useful resource for researchers who propose new datasets to assess the suitability and quality of their data to evaluate various phenomena of interest, as well as those who propose novel NLU approaches, to further understand the implications of their improvements with respect to their model's acquired capabilities.

Keywords: Natural Language Understanding; Deep learning; Machine reading comprehension; Textual entailment; Dataset artefacts

1. Introduction

Research in areas that require reasoning over and understanding unstructured, natural language text, is advancing at an unprecedented rate. Novel neural architectures (Vaswani *et al*. 2017) enable efficient unsupervised training on large corpora to obtain expressive contextualised word and sentence representations as a basis for a multitude of downstream NLP tasks (Devlin *et al*. 2019). They are further fine-tuned on task-specific, large-scale datasets ((Bowman *et al*. 2015; Rajpurkar *et al*. 2016); Williams, Nangia, and Bowman 2018) which provide sufficient examples to optimise large neural models that are capable of outperforming human-established baselines on multiple Natural Language Understanding (NLU) benchmarks (Raffel *et al*. 2020; Lan *et al*. 2020). This seemingly superb performance is used as a justification to accredit those models various NLU capabilities, such as numeric reasoning (Dua *et al*. 2019b), understanding the temporality of events (Zhou *et al*. 2019) or integrating information from multiple sources (Yang *et al*. 2018).

Recent work, however, casts doubts on the capabilities obtained by models optimised on these data. Specifically, they may contain exploitable superficial cues, for example, the most frequent answer to questions of the type *'How many. . .'* is *'2'* in a popular numeric reasoning dataset (Gardner *et al*. 2020) or the occurrence of the word *'no'* is correlated with non-entailment in large Recognising Textual Entailment (RTE) datasets (Gururangan *et al*. 2018). Models are evaluated following the usual machine learning protocol, where a random subset of the dataset is withheld

for evaluation under a performance metric. Because the subset is drawn *randomly*, these correlations exist in the evaluation data as well and models that learn to rely on them obtain a high score. While exploiting correlations is in itself not a problem, it becomes an issue when they are *spurious*, that is, they are artefacts of the collected data rather than representative of the underlying task. As an example, answering '2' to the question 'How many. . .' is evidently not representative of the task of numeric reasoning.

A number of publications identifies weaknesses of training and evaluation data and whether optimised models inherit them. Meanwhile, others design novel evaluation methodologies that are less prone to the limitations discussed earlier and therefore establish more realistic estimates of various NLU capabilities of state-of-the-art models. Yet others propose improved model optimisation practices which aim to ignore flaws in training data. The work by McCoy, Pavlick and Linzen (2019) serves as an example for the coherence of these research directions: first, they show that in crowdsourced RTE datasets, specific syntactic constructs are correlated with an expected class. They show that optimised models rely on this correlation, by evaluating them on valid counterexamples where this correlation does not hold. Later, they show that increasing the syntactic diversity of training data helps to alleviate these limitations (Min *et al*. 2020).

In this paper, we present a structured survey of this growing body of literature. We survey 121 papers for methods that reveal and overcome weaknesses in data and models and categorise them accordingly. We draw connections between different categories, report the main findings, discuss arising trends and cross-cutting themes, and outline open research questions and possible future directions. Specifically, we aim to answer the following questions:

(1) Which NLU tasks and corresponding datasets have been investigated for weaknesses?
(2) Which types of weaknesses have been reported in models and their training and evaluation data?
(3) What types of methods have been proposed to detect and quantify those weaknesses and measure their impact on model performance, and what methods have been proposed to overcome them?
(4) How have the proposed methods impacted the creation and publication of novel datasets?

The paper is organised as follows: we first describe the data collection methodology and describe the collected literature body. We then synthesise the weaknesses that have been identified in this body and categorise the methods used to reveal those. We highlight the impact of those methods on the creation of new resources and conclude with a discussion of open research questions as well as possible future research directions for evaluating and improving the NLU capabilities of NLP models.

2. Methodology

To answer the first three questions, we collect a literature body using the 'snowballing' technique. Specifically, we initialise the set of surveyed papers with Tsuchiya (2018), Gururangan *et al*. (2018), Poliak *et al*. (2018) and Jia and Liang (2017), because their impact helped to motivate further studies and shape the research field. For each paper in the set, we follow its citations and any work that has cited it according to Google Scholar and include papers that describe methods and/or their applications to report any of: (1) qualitative and quantitative investigation of flaws in training and/or test data and the impact on models optimised/evaluated thereon; (2) systematic issues with task formulations and/or data collection methods; (3) analysis of specific linguistic and reasoning phenomena in data and/or models' performance on them or (4) proposed improvements in order to overcome data-specific or model-specific issues, related to the phenomena and flaws described earlier. We exclude a paper if its target task is not concerning NLU and was published before the

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

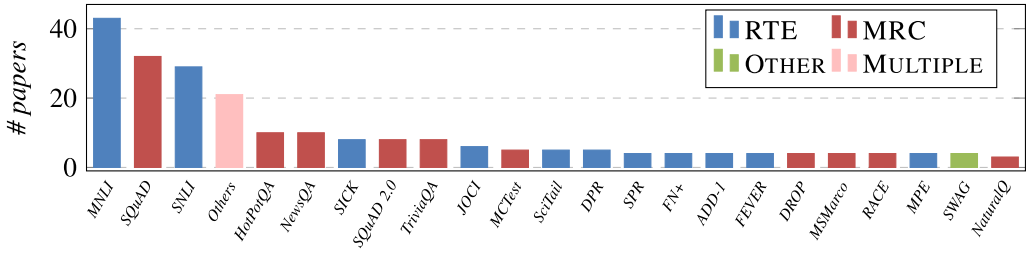

Figure 1. Bar chart with RTE, MRC and other datasets that were investigated by at least three surveyed papers. Datasets investigated once or twice are summarised with 'Multiple'. Full statistics can be observed in the Appendix.

year 2014, or the language of the investigated data is not English. We set 2014 as lower boundary, because it precedes the publication of most large-scale crowdsourced datasets that require NLU.

With this approach, we obtain a total of 121 papers (as of 17 October 2020) from the years 2014 to 2017 (8), 2018 (18), 2019 (42) and 2020 (53). Almost two-thirds (76) of the papers were published in venues hosted by the the Association for Computational Linguistics. The remaining papers were published in other venues (eight in AAAI, four in LREC, three in ICLR, two in ICML and COLING, respectively, five other) or are available as an arXiv preprint (21). The papers were examined by the first author; for each paper the target task and dataset(s), the method applied and the result of the application was extracted and categorised.

To answer the fourth question, we selected those publications introducing any of the datasets that were mentioned by at least one paper in the pool of surveyed papers and extended that collection by additional state-of-the-art NLU dataset resource papers (for detailed inclusion and exclusion criteria, see Appendix A). This approach yielded a corpus of 91 papers that introduce 95 distinct datasets. For those papers, we examine whether any of the previously collected methods were applied to report spurious correlations or whether the dataset was adversarially pruned against some model.

Although related, we deliberately do not include work that introduces adversarial attacks on NLP systems or discusses their fairness, as these are out of scope of this survey. For an overview thereof, we refer the interested reader to respective surveys conducted by Zhang *et al.* (2020b) or Xu *et al.* (2020) for the former, and by Mehrabi *et al.* (2021) for the latter. Furthermore, we do not include works that concern wider technical issues, such as performance variance due to different software environments (Crane and Cheriton 2018) or stochastic instability (Dodge *et al.* 2020).

3. Investigated tasks and datasets

We report the tasks and the corresponding datasets that we have investigated. We supply a full list of these investigated datasets and the type(s) of method(s) applied in Appendix B. Figure 1 depicts all investigated datasets as a bar chart. The distribution roughly follows the popularity of the investigated resources themselves, with the papers presenting MNLI, SQuAD and SNLI being the three most cited among the investigated datasets.[a]

Almost half of the surveyed papers (57) are focused on the RTE task, where the goal is to decide, for a pair of natural language sentences (premise and hypothesis), whether given the premise the hypothesis is true (*Entailment*), certainly false (*Contradiction*), or whether the hypothesis might be true, but there is not enough information to determine that (*Neutral*) (Dagan *et al.* 2013).

Many of the papers analyse the Machine Reading Comprehension (MRC) task (50 papers), a special case of Question Answering (QA) which concerns finding the correct answer to a question over a passage of text. Note that the tasks are related: answering a question can be framed as finding an answer that is entailed by the question and the provided context (Demszky, Guu and

[a] As of November 2021, according to Google Scholar.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Liang 2018). Inversely, determining whether a hypothesis is true given a premise can be framed as question answering.

Other tasks (eight papers) involve finding the most plausible cause or effect for a short prompt among two alternatives (Roemmele, Bejan and Gordon 2011), fact verification (Thorne *et al*. 2018) and argument reasoning (Habernal *et al*. 2018). Seven papers investigated multiple tasks. Note, that weaknesses reported in the surveyed papers were also reported on data and models representing tasks typically not associated with NLU, such as sentiment analysis (Ko *et al*. 2020a). To keep this paper within the scope we set out, we do not include them in the following discussions.

In general, 18 RTE and 37 MRC datasets were analysed or used at least once; we attribute this difference in number to the existence of various MRC datasets and the tendency of performing multi-dataset analyses in papers that investigate MRC datasets (Kaushik and Lipton 2018; Si *et al*. 2019; Sugawara *et al*. 2020). SQuAD (Rajpurkar *et al*. 2016) for MRC and MNLI (Williams *et al*. 2018) and SNLI (Bowman *et al*. 2015) for RTE are the most utilised datasets in the surveyed literature (with 32, 43 and 29 papers investigating or using them).

4. Identified weaknesses in NLU data and models

In this section, we present the types of weaknesses that have been reported in the surveyed literature. State-of-the-art approaches to solve the investigated tasks are predominantly data-driven. We distinguish between issues identified in their training and evaluation data on the one hand, and the extent to which these issues affect the trained models on the other hand.

4.1 Weaknesses in data

Spurious correlations: Correlations between input data and the expected prediction are 'spurious' if there exists no causal relation between them with regard to the underlying task but rather they are an artefact of a specific dataset. They are also referred to as '(annotation) artefacts' (Gururangan *et al*. 2018) or '(dataset) biases' (He *et al*. 2019) in literature.

In span extraction tasks, where the task is to predict a continuous span of token in text, as is the case with MRC, question and passage wording, as well as the position of the answer span in the passage are indicative of the expected answer for various datasets (Rychalska *et al*. 2018a; Kaushik and Lipton 2018) such that models can solve examples correctly even without being exposed to either the question or the passage. In the ROCStories dataset (Mostafazadeh *et al*. 2016) where the task is to choose the most plausible ending to a story, the writing style of the expected ending differs from the alternatives (Schwartz *et al*. 2017). This difference is noticeable even by humans (Cai, Tu and Gimpel 2017).

For sentence pair classification tasks, such as RTE, Poliak *et al*. (2018) and Gururangan *et al*. (2018) showed that certain *n*-grams, lexical and grammatical constructs in the hypothesis as well as its length correlate with the expected label for a multitude of RTE datasets. For example, the word 'no' in the premise occurs more often with the label *Contradiction* than with the label *Entailment* in the SNLI and MNLI datasets. This correlation is *spurious* because although true for the datasets, the appearance of the word 'no' in the premise is not indicative of contradiction. For example, 'No cats are in the room.' entails 'No cats are under the bed.'. Similarly, McCoy *et al*. (2019) showed that lexical features like word overlap and common subsequences between the hypothesis and premise are highly predictive of the entailment label in the MNLI dataset. These correlations, too, are artefacts of the data collection method rather than robust indicators of entailment – for example 'I *almost* went to Vienna.' does not entail 'I went to Vienna' despite a high lexical overlap.

Beyond RTE, the choices in the COPA dataset (Roemmele *et al*. 2011) where the task is to finish a given passage (similar to ROCStories) and ARCT (Habernal *et al*. 2018) where the task is to select

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Passage 1: Marietta Air Force Station *Marietta Air Force Station (ADC ID: M-111, NORAD ID: Z-111) is a closed United States Air Force General Surveillance Radar station. It is located 2.1 mi northeast of Smyrna, Georgia. It was closed in 1968.*	**Passage 2: Smyrna, Georgia** *Smyrna is a city northwest of the neighborhoods of Atlanta. It is in the inner ring of the Atlanta Metropolitan Area. As of the 2010 census, the city had a population of **51,271**. The U.S. Census Bureau estimated the population in 2013 to be 53,438. [. . .]*
Passages 3-10: *[...]*	
Question: *What is the 2010 population of the city 2.1 miles southwest of Marietta Air Force Station?*	

Figure 2. Example for a dataset artefact where the requirement to synthesise information from 2 out of 10 accompanying passages can be circumvented by exploiting simple word co-occurrence between question and answer sentence.

whether a statement warrants a claim contain words that correlate with the expected prediction (Kavumba *et al*. 2019; Niven and Kao 2019).

Other data quality issues: Pavlick and Kwiatkowski (2019) argue that when training data are annotated using crowdsourcing, a fixed label representing the ground truth, usually obtained by majority vote between annotators, is not representative of the uncertainty, which can be important to indicate the complexity of an example. Sometimes, data annotations are factually wrong (Pugaliya *et al*. 2019; Schlegel *et al*. 2020), for example, due to limitations of the annotation protocol, such as when only one ground truth label is expected but the data contain multiple plausible alternatives. In 'multi-hop' datasets, such as HotpotQA and WikiHop where the task is to find an answer after aggregating evidence across multiple documents, this process can be circumvented in the case of examples where the location of the final answer is cued by the question (Min *et al*. 2019). For example, consider Figure 2: while initially this looks like a complex question that requires spatial reasoning over multiple documents, the keyword combination '2010' and 'population' in the question is unique to the answer sentence across all 10 documents, allowing to find the answer to the question without ever reading Passage 1. The initially complex question can be substituted by the much easier question 'What is the 2010 population?' which does not require any reasoning and has a unique answer that coincides with the expected answer to the original question. This is especially true for the multiple-choice task formulation, as the correct answer can often be 'guessed' by excluding implausible alternatives (Chen and Durrett 2019), for example, by matching the interrogative pronoun with the corresponding lexical answer type. Sugawara *et al*. (2018) show that multiple MRC benchmarks contain numerous questions that are easy to answer, as they do require little comprehension or inference skills and can be solved by looking at the first few tokens of the question. This property appears ubiquitous among multiple datasets (Longpre, Lu and DuBois 2021). Finally, Rudinger, May and Van Durme (2017) show the presence of gender and racial stereotypes in crowdsourced RTE datasets.

There are multiple reasons for data quality issues, among them the *carelessness* of the annotators, usually hired via a crowdsourcing platform (Brühlmann *et al*. 2020). This also provides a possible explanation to the existence of dataset artefacts: as annotators are paid per annotation, they often adapt simple strategies to maximise the output. For example, deriving a contradicting premise by simply negating the hypothesis might lead to the spurious correlation discussed before. Thus, it is important to establish the quality of data during and after collection, by filtering low-quality examples, an increasingly employed practice, as we discuss in Section 5.1.

In any case, these data quality issues diminish the explanatory power of observations about models evaluated on these data: the presence of cues casts doubts on the requirements of various NLU capabilities, if a simpler model can perform reasonably well by exploiting these cues. The situation is similar, when expected answers are factually wrong.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

4.2 Model weaknesses

In this section, we discuss whether data-driven approaches to NLU are in fact affected by these data quality issues. We discuss multiple works that reveal their *dependence on dataset-specific artefacts*. This is further evidenced by their *poor generalisation* on data that stem from a different distribution than their training data, suggesting that they in fact overfit to the patterns inherent to different datasets rather than reliably learning the underlying task. A suspected reason for this is that state-of-the-art NLP approaches rely increasingly on *no-assumption architectures* – little to no expert knowledge is encoded into the models a priori and all necessary information is expected to be derived from the data.

Dependence on dataset-specific artefacts: Given the data-related issues discussed earlier, it is worth knowing whether models optimised on this data actually inherit them. In fact, multiple studies confirm this hypothesis (McCoy *et al*. 2019; Niven and Kao 2019; Kavumba *et al*. 2019). This is usually evidenced by the poor performance of models evaluated on a balanced version of the data where the spurious correlations have been balanced. To illustrate it by means of the previous example, a balanced set would have equally many instances with the word 'no' appearing with the label *Contradiction* and *Entailment*. Note, that the act of balancing the data alters the underlying generative process; thus, this type of evaluation is performed on data outside of the training data distribution; therefore, the same considerations regarding poor out-of-distribution generalisation, discussed in more detail in the following section, apply.

Neural models tend to disregard syntactic structure (Rychalska *et al*. 2018b,a) and important words (Mudrakarta *et al*. 2018), making them *insensitive* towards small but potentially meaningful perturbations in inputs. This results in MRC models that are negatively impacted by the presence of lexically similar but semantically irrelevant 'distractor sentences' (Jia and Liang 2017; Jiang and Bansal 2019), give inconsistent answers to semantically equivalent input (Ribeiro, Singh and Guestrin 2018) or fail to distinguish between semantically different inputs with similar surface form (Gardner *et al*. 2020; Welbl *et al*. 2020). For RTE, they may disregard the composition of the sentence pairs (Nie, Wang and Bansal 2019).

Poor generalisation outside of training distribution: Mediocre performance when evaluated on RTE (Glockner, Shwartz and Goldberg 2018; Naik *et al*. 2018; Yanaka *et al*. 2019b) and MRC data (Talmor and Berant 2019; Dua *et al*. 2019a) that stems from a different generative process than the training data (leading to out-of-distribution examples) reinforces the fact that models pick up spurious correlations that do not hold between different datasets, as outlined earlier. Limited out-of-distribution generalisation capabilities of state-of-the-art models suggest that they are 'lazy learners': when possible, they infer simple decision strategies from training data that are not representative of the corresponding task, instead of learning the necessary capabilities to perform inference. Nonetheless, recent work shows that the self-supervised pre-training of transformer-based language models allows them to adapt to the new distribution from few examples (Brown *et al*. 2020; Schick and Schütze 2021).

No-assumption architectures: Note that these weaknesses arise because state-of-the-art end-to-end architectures[b] (Bahdanau, Cho and Bengio 2015), such as the transformer (Vaswani *et al*. 2017), are designed with minimal assumptions. As little as possible prior knowledge is encoded into the model architecture – all necessary information is expected to be inferred from the (pre-)training data. The optimisation objectives reflect this assumption as well: beyond the loss function accounting for the error in prediction, hardly any regularisation is used. As a consequence, there is no incentive for models to distinguish between spurious and reliable correlations, so they follow the strongest signal present in data. In fact, one of the main themes discussed in Section 5.3 is to inject additional knowledge, for example, in the form of more training data or

[b]Note that we refer to the neural network architecture of a model as 'architecture', for example, BiDAF (Seo *et al*. 2017), while we refer to a (statistical) model of a certain architecture that was optimised on a specific training set simply as 'model'.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

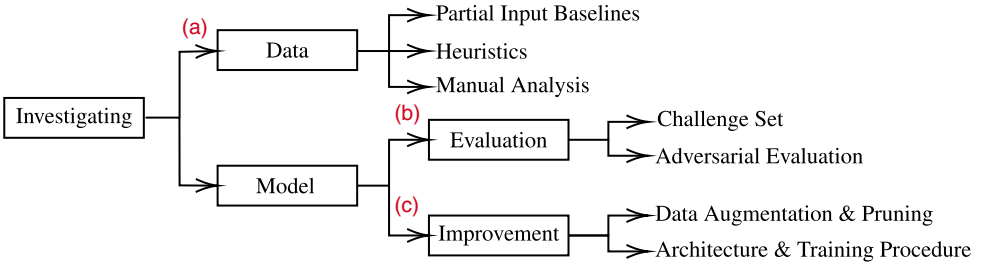

Figure 3. Taxonomy of investigated methods. Labels (a), (b) and (c) correspond to the coarse grouping discussed in Section 5.

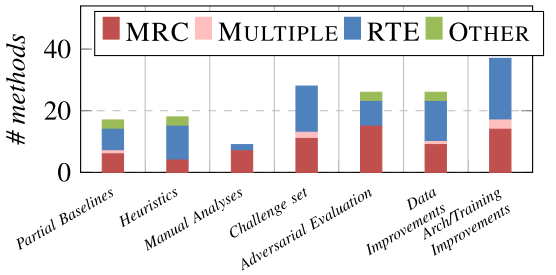

Figure 4. Number of methods per category split by task. As multiple papers report more than one method, the maximum (160) does not add up to the number of surveyed papers (121).

heavier regularisation, as a countermeasure, in order to make the optimised model rely less on potentially biased data. For example, models that operate over syntax trees rather than sequences tend to be less prone to syntactic biases (McCoy *et al.* 2019).

5. Categorisation of methods that reveal and overcome weaknesses in NLU

In the following section, we categorise the methodologies collected from the surveyed papers, briefly describe the categories and exemplify them by referring to respective papers. On a high level, we distinguish between methods that: (a) reveal systematic issues with existing training and evaluation data, such as the spurious correlations mentioned earlier, (b) investigate whether they translate to models optimised on these data with regard to acquired inference and reasoning capabilities and (c) propose architectural and training procedure improvements in order to alleviate the issues and improve the robustness of the investigated models. A schematic overview of the taxonomy of the categories is shown in Figure 3. The quantitative results of the categorisation are shown in Figure 4.

5.1 Data-investigating methods

Methods in this category analyse flaws in data such as cues in input that are predictive of the output (Gururangan *et al.* 2018). As training and evaluation data from state-of-the-art NLU datasets are assumed to be drawn from the same distribution, models that were fitted on those cues achieve high performance in the evaluation set, without being tested on the required inference capabilities. Furthermore, methods that investigate the evaluation data in order to better understand the assessed capabilities (Chen, Bolton and Manning 2016) fall under this category as well. In the analysed body of work, we identified the types of methods discussed in the following paragraphs. In Table 1, we summarise them with their corresponding investigation goal.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Table 1. Summary of data-investigating methods with the corresponding research questions as described in Section 5.1

Method	Task	Target weakness	Pursued research question
Partial input baselines	Any with multiple input parts	Spurious correlations	Are all parts of the input required for the prediction?
Data ablation	Any	Data quality	Is the capability represented by the removed data necessary to solve the dataset?
Architectural constraint	Any	Data quality	Is the capability restricted by the constraint necessary to solve the dataset?
Heuristics	Classification	Spurious correlations	Which features that correlate with the expected label are spurious?
Manual analysis	Any	Data quality	Does the data represent the challenges of the underlying task?

Partial Baselines are employed in order to verify that all input provided by the task is actually required to make the right prediction (e.g., both question and passage for MRC, and premise and hypothesis for RTE). If a classifier trained on partial input performs significantly better than a random guessing baseline, it stands to reason that the omitted parts of the input are not required to solve the task. On the one hand, this implies that the input used to optimise the classifier might exhibit cues that simplify the task. On the other hand, if the omitted data represent a specific capability, the conclusion is that this capability is not evaluated by the dataset, a practice we refer to as *Data Ablation*. Examples for the former include training classifiers that perform much better than the random guess baseline on hypotheses only for the task of RTE (Gururangan *et al.* 2018; Poliak *et al.* 2018) and on passages only for MRC (Kaushik and Lipton 2018).[c] For the latter, Sugawara *et al.* (2020) drop words that are required to perform certain comprehension abilities (e.g., dropping pronouns to evaluate pronominal coreference resolution capabilities) and reach performance comparable to that of a model that is trained on the full input on a variety of MRC datasets. Nie *et al.* (2019) reach near state-of-the-art performance on RTE tasks when shuffling words in premise and hypothesis, showing that understanding the compositional nature of language is not required by these datasets. A large share of work in this area concentrates on evaluating datasets with regard to the requirement to perform 'multi-hop' reasoning (Min *et al.* 2019; Chen and Durrett 2019; Jiang and Bansal 2019; Trivedi *et al.* 2020) by measuring the performance of a partial baseline that exhibits an *Architectural Constraint* to perform single-hop reasoning (e.g., by processing input sentences independently).

Insights from partial baseline methods bear negative predictive power only – their failure does not necessarily entail that the data are free of cues, as they can exist in different parts of the input. As an example, consider an MRC dataset, where the three words before and after the answer span are appended to the question. Partial baselines would not be able to pick up this cue, because it can only be exploited by considering both question and passage. Feng, Wallace and Boyd-Graber (2019) show realistic examples of this phenomenon in published datasets. Furthermore, above-chance performance of partial baselines merely hints at spurious correlations in the data and suggests that models learn to exploit them; it does not reveal their precise nature.

Heuristics and Correlations are used to unveil the nature of cues and spurious correlations between input and expected output. For sentence pair classification tasks, modelling the co-occurrence of words or n-grams with the expected prediction label by means of Pointwise Mutual

[c] In some cases, they even match or surpass the performance of the reference full-input model.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Information (Gururangan *et al*. 2018) or conditional probability (Poliak *et al*. 2018; Tan *et al*. 2019) shows the likelihood of an expression being predictive of a label. Measuring coverage (Niven and Kao 2019) further indicates what proportion of the dataset is affected by this correlation. Manually inspecting these correlations can help to identify whether they can simplify the task. For example, if an expression perfectly correlates with a label and covers some subset of the data, then this subset can be solved correctly by relying on the appearance of this expression alone. They are further spurious, if they are not indicative of the underlying task but rather artefacts of the dataset.

These exploratory methods require no apriori assumptions about the kind of bias they can reveal. Other methods require more input, such as qualitative data analysis and identification of syntactic (McCoy *et al*. 2019) and lexical (Liu *et al*. 2020b) patterns that correlate with the expected label. Furthermore, Nie *et al*. (2019) use the confidence of a logistic regression model optimised on lexical features to predict the wrong label to rank data by their requirements to perform comprehension beyond lexical matching.

It is worth highlighting that there is comparatively little work analysing MRC data (4 out of 18 surveyed methods) with regard to spurious correlations. We attribute this to the fact that it is hard to conceptualise the correlations of input and expected output for MRC beyond very coarse heuristics such as sentence position (Si *et al*. 2020) or lexical overlap (Sugawara *et al*. 2018), as the input is a whole paragraph and a question and the expected output is typically a span anywhere in the paragraph. Furthermore, the prediction labels (paragraph indices for answer spans or the number of the chosen alternative for multiple-choice type of questions) do not bear any semantic meaning, so correlation between input and predicted raw output such as those discussed earlier can only unveil positional bias. For RTE, in contrast, the input consists of two sentences and the expected output is one of three fixed class labels that carry the same semantics regardless of the input; therefore, possible correlations are easier to unveil.

Manual Analyses are performed to qualitatively analyse the data, if automated approaches as those mentioned earlier are unsuitable due to the complexity of the phenomena of interest or the output space discussed earlier. We posit that this is the reason why most methods in this category concern analysing MRC data (seven out of nine surveyed methods). Qualitative annotation frameworks were proposed to investigate the presence of linguistic features (Schlegel *et al*. 2020) and cognitive skills required for reading comprehension (Sugawara, Yokono and Aizawa 2017a).

5.2 Model-investigating methods

Rather than analysing data, approaches described in this section directly evaluate models in terms of their inference capabilities with respect to various phenomena of interest. Released evaluation resources are summarised in Table 2.

Challenge Sets make for an increasingly popular way to assess various capabilities of optimised models. Challenge sets feature a collection of (typically synthetically generated) examples that exhibit a specific phenomenon of interest. Bad performance on the challenge set indicates that the model has failed to obtain the capability to process the phenomenon correctly. Similar to partial baselines, a good result does not necessarily warrant the opposite, unless guarantees can be made that the challenge set is perfectly representative of the investigated phenomenon. Naik *et al*. (2018) automatically generate RTE evaluation data based on an analysis of observed state-of-the-art model error patterns, introducing the term 'stress-test'. Challenge sets have since been proposed to evaluate RTE models with regard to the acquisition of linguistic capabilities such as monotonicity (Yanaka *et al*. 2019a), lexical inference (Glockner *et al*. 2018), logic (Richardson *et al*. 2019) and understanding language compositionality (Nie *et al*. 2019). With respect to MRC, we note that there are few (11) challenge sets concerning rather broad categories such as prediction consistency (Ribeiro *et al*. 2019; Gardner *et al*. 2020), acquired knowledge (Richardson and Sabharwal 2020) or transfer to different datasets (Dua *et al*. 2019a; Miller *et al*. 2020).

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Table 2. Proposed adversarial and challenge evaluation sets with their target phenomenon, grouped by task and, where appropriate, with original resource name. The last column 'OOD'indicates, whether the authors acknowledge and discount for the distribution shift between training and challenge set data (Y), they do not (N), whether performance under the distribution shift is part of the research question (P), whether an informal argument (I) is provided or whether it is not applicable (-)

Task	Challenge set	Target weakness, phenomenon or capability	OOD
MRC	AddSent (Jia and Liang 2017)	Dependence on word overlap between question and answer sentence	P
	AddDoc (Jiang and Bansal 2019)	Dependence on word overlap between question and answer sentence to circumvent 'multi-hop' reasoning	P
	Ribeiro, Guestrin and Singh (2019)	Consistency on semantically equivalent input	N
	Nakanishi, Kobayashi and Hayashi (2018)	Answering unanswerable questions for MRC	N
	Tang, Ng and Tung (2021)	Answering decomposed "multi-hop" questions	P
	(Trivedi *et al*. 2020)	Identifying whether presented facts are sufficient to justify the answer to a "multi-hop" question	–
	ContrastSet (Gardner *et al*. 2020)	Sensitivity to meaningful input perturbations	P
	(Miller *et al*. 2020)	Performance under domain shift	P
RTE	HANS (McCoy *et al*. 2019)	Dependence on syntax and word overlap	P
	Salvatore, Finger and Hirata (2019)	Understanding negation, coordination, quantifiers, definite descriptions, comparatives and counting	N
	(Richardson *et al*. 2019)	Understanding the capabilities from (Salvatore *et al*. 2019), conditionals and monotonicity	Y
	ImpPres (Jeretic *et al*. 2020)	Understanding implicature and presupposition	N
	Goodwin, Sinha and O'Donnell (2020)	Systematic generalisation of the understanding of compositionality in an artificial language	–
	CompSens (Nie *et al*. 2019)	Understanding the compositionality of sentences	–
	BreakingNLI (Glockner *et al*. 2018)	Understanding lexical entailments	N
	TaxiNLI (Joshi *et al*. 2020)	Performing various reasoning capabilities	–
	(Rozen *et al*. 2019)	Dative alteration and numerical ordering	Y/P
	HypOnly (Gururangan *et al*. 2018)	Dependence on spurious artefacts in hypothesis	–
	MED (Yanaka *et al*. 2019a)	Understanding monotonicity reasoning	I
	StressTest (Naik *et al*. 2018)	Dependence on lexical similarity, sentence length and correct spelling; understanding numerals, negations, antonyms	N
	NerChanged (Mitra, Shrivastava, and Baral 2020)	Different named entities in identical situations	N
	RolesSwitched (Mitra *et al*. 2020)	Asymetry of verb predicates	N
	(Nie *et al*. 2019)	Understanding the compositionality of language	–
	Kaushik, Hovy, and Lipton (2020)	Consistency on counterfactual examples	P
	TeachYourAI (Talmor *et al*. 2020)	Reasoning with implicit knowledge	–
MCQA	(Richardson and Sabharwal 2020)	Understanding word senses and definitions	–

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Table 2. Continued

Task	Challenge set	Target weakness, phenomenon or capability	OOD
Fact Checking	FEVER-B (Schuster *et al*. 2019)	Dependence on lexical surface form	P
Argument Reasoning	ARCT2 (Niven and Kao 2019)	Dependence on spurious lexical trigger words	P
Common-sense Reasoning	B-COPA (Kavumba *et al*. 2019)	Dependence on spurious lexical trigger words	P

Notably, these challenge sets are well suited to evaluate the investigated capabilities, because they perform a form of *out-of-distribution* evaluation. Since the evaluation data stem from a different (artificial) generative process than the crowdsourced training data, possible decision rules based on cues are more likely to fail. The drawback of this, however, is that in this way the challenge sets evaluate both the investigated capability and the performance under distribution shift. Liu, Schwartz and Smith (2019a) show that for some of the challenge sets, after fine-tuning ('inoculating') on small portions of it, the challenge set performance increases, without sacrificing the performance on the original data. However, Rozen *et al*. (2019) show that good performance after fine-tuning cannot be taken as evidence of the model learning the phenomenon of interest – rather the model adapts to the challenge-set-specific distribution and fails to capture the general notion of interest. This is indicated by low performance when evaluating on challenge sets that stem from a different generative process but focus on the same phenomenon. These results suggest that the 'inoculation' methodology is of limited suitability to disentangle the effects of domain shift from evaluating the capability to process the investigated phenomenon.

Furthermore, a line of work proposes to evaluate the systematic generalisation capabilities of RTE models (Geiger *et al*. 2019; Geiger, Richardson and Potts 2020; Goodwin *et al*. 2020), concretely the capability to infer and understand compositional rules that underlie natural language. These studies concern mostly artificial languages, however.

Adversarial Evaluation introduces evaluation data that were generated with the aim to 'fool' models. Szegedy *et al*. (2014) define 'adversarial examples' as (humanly) imperceptible perturbations to images that cause a significant drop in the prediction performance of neural models. Similarly for NLP, we refer to data as 'adversarial' if it is designed to minimise prediction performance for automated approaches, while not impacting the human baseline. Thus, patterns in the behaviour of a certain model or a range of models of certain architectures serve as a starting point for the development of adversarial evaluation methods. This is different from challenge sets, which focus on phenomena of interest without assumptions about the evaluated model. Notably, these two terms are sometimes used interchangeably in literature.

Adversarial methods are used to show that models rely on superficial, dataset-specific cues, as discussed in Section 4.2. This is typically done by creating a balanced version of the evaluation data, where the previously identified spurious correlations present in training data do not hold anymore (McCoy *et al*. 2019; Kavumba *et al*. 2019; Niven and Kao 2019), or by applying semantic preserving perturbations to the input (Jia and Liang 2017; Ribeiro *et al*. 2018). Note that this is yet another method that alters the distribution of the evaluation data with respect to the training data.

Adversarial techniques are further used to understand model behaviour (Sanchez, Mitchell and Riedel 2018), such as identifying training examples (Han, Wallace and Tsvetkov 2020) or neuron activations (Mu and Andreas 2020) that contribute to a certain prediction. Among those we highlight the work by Wallace *et al*. (2019) who showed that malicious adversaries generated against a target model tend to be universal for a whole range of neural architectures.

5.3 Model-improving methods

Here, we report methods that improve the robustness of models against adversarial and out-of-distribution evaluation, by either modifying the training data or making adjustments to model

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

architecture or the training procedure. We group the methods by their conceptual approach and present them together with their applications in Table 3. In line with the literature (Wang and Bansal 2018; Jia *et al.* 2019), we call a model 'robust' against a method that alters the underlying distribution of the evaluation data (hence making it substantially different from the training data) through for example, adversarial or challenge sets, if the out-of-distribution performance of the model is similar to that on the original evaluation set. They have become increasingly popular: 30%, 35% and 51% of the surveyed methods published in the years 2018, 2019 and 2020, respectively, fall into this category (and none before 2018). We attribute this to the public availability of evaluation resources discussed in Section 5.2 as they facilitate the rapid prototyping and testing of these methods.

Data Augmentation and Pruning combat the issues arising from low-bias architecture by injecting the required knowledge, in the form of (usually synthetically generated) data, during training. There is ample evidence that augmenting training data with examples featuring a specific phenomenon increases the performance on a challenge set evaluating that phenomenon (Wang *et al.* 2018; Jiang and Bansal 2019; Zhou and Bansal 2020)—for example, Yanaka *et al.* (2019b) propose an automatically constructed dataset as an additional training resource to improve monotonicity reasoning capabilities in RTE. As these augmentations come at the cost of lower performance on the original evaluation data, Maharana and Bansal (2020) propose a framework to combine different augmentation techniques such that the performance on both is optimised.

More interesting are approaches that augment data without focussing on a specific phenomenon. By increasing data diversity, better performance under adversarial evaluation can be achieved (Talmor and Berant 2019; Tu *et al.* 2020). Similarly, augmenting training data in a *meaningful* way, for example, with counterexamples, by asking crowdworkers to apply perturbations that change the expected label (Kaushik *et al.* 2020; Khashabi *et al.* 2020), helps models to achieve better robustness beyond the training set distribution.

An alternative direction is to increase data *quality*, by removing data points that exhibit spurious correlations. After measuring the correlations with methods discussed in Section 5.1, those training examples exhibiting strong correlations can be removed. The AFLITE algorithm (Sakaguchi *et al.* 2020) combines both of these steps by assuming that a linear correlation between embeddings of inputs and prediction labels is indicative of biased data points. This is an extension of the *Adversarial Filtering* algorithm (Zellers *et al.* 2018), whereby multiple-choice alternatives are automatically generated until a target model can no longer distinguish between human-written (correct) and automatically generated (wrong) options.

A noteworthy trend is the application of *adversarial data generation* against a target model that is employed during the construction of a new dataset. In crowdsourcing, humans act as adversary generators and an entry is accepted only if it triggers a wrong prediction by a trained target model (Dua *et al.* 2019b; Nie *et al.* 2020). Mishra *et al.* (2020) combine both directions in an interface which aims to assist researchers who publish new datasets with different visualisation, filtering, and pruning techniques.

Architecture and Training Procedure Improvements deviate from the idea of data augmentation and seek to train robust models from potentially biased data. Adversarial techniques (Goodfellow *et al.* 2014), in which a generator of adversarial training examples (such as those discussed in Section 5.2, e.g., perturbing the input) is trained jointly with the discriminative model that is later used for inference, have been applied to different NLU tasks (Stacey *et al.* 2020; Welbl *et al.* 2020).

Specific knowledge about the type of bias present in data can be used to discourage a model from learning from it. For example, good performance (as indicated by a small loss) of a partial input classifier is interpreted as an indication that data points exhibit spurious correlations. This information can be used to train an 'unbiased' classifier jointly (Clark et al. 2019b; He *et al.* 2019; Belinkov *et al.* 2019). Alternatively, their contribution to the overall optimisation objective can be rescaled (Schuster *et al.* 2019; Zhang *et al.* 2020b; Mehrabi *et al.* 2021). The intuition behind these approaches is similar to *Adversarial Filtering* which is mentioned earlier: the contribution of biased data to the overall training is reduced. For lexical biases, such as cue words, Utama

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Table 3. Categorisation of methods that have been proposed to overcome weaknesses in models and data. To indicate that a method was applied to improve performance on a challenge set, we specify the challenge set name as presented in Table 2

Approach	*Description* → Applications
Data Augmentation	*Uses additional training data to improve performance on a phenomenon or to combat a model weakness*
	→ Counterfactual augmentation for RTE and MRC ((Ko *et al*. 2020a); Khashabi, Khot and Sabharwal 2020; Asai and Hajishirzi 2020), adversarially generated training data for MRC (Jiang and Bansal 2019; Wang and Bansal 2018; Yang *et al*. 2020), Monotonicity reasoning for RTE (Yanaka *et al*. 2019b)
Adversarial Filtering	*Minimises dataset artefacts by removing/replacing data points that can be predicted with high confidence during multiple cross-validation runs*
	→ Removal of spurious correlations in commonsense reasoning datasets (Zellers *et al*. 2018; 2019; Sakaguchi *et al*. 2020; Bras *et al*. 2020)
Humans as adversaries	*Ground truth annotations from crowdworkers are only approved if an optimised model cannot predict them*
	→ Applied for RTE (Nie *et al*. 2020), MRC (Dua *et al*. 2019b) and MCQA (Chen et al. 2019) datasets
Bias Ensembling	*Trains a robust model with an artificially biased model; this discourages the robust model to learn biases picked up by biased model* → Answer position Bias in MRC (Ko *et al*. 2020b), AddSent (Clark, Yatskar and Zettlemoyer 2019b); synthetic data, HANS and StressTest (He, Zha and Wang 2019; Karimi Mahabadi, Belinkov and Henderson 2020; Zhou and Bansal 2020), HypOnly and transfer learning between RTE datasets (Belinkov *et al*. 2019)
Downweighting	*Scales down the contribution of biased data points (e.g., identified by partial baseline methods) to the overall loss minimising objective of the training set* → FeVeR-B (Schuster *et al*. 2019), HypOnly, HANS and transfer learning between RTE datasets (Zhang *et al*. 2019b; Karimi Mahabadi *et al*. 2020; Utama, Moosavi and Gurevych 2020), . . .
Example Forgetting	*Identifies examples that are misclassified during training as 'hard' examples; hard examples are used for additional fine-tuning* → HANS (Yaghoobzadeh *et al*. 2019)
Regularisation with expert knowledge	*Uses regularisation terms to encode expert domain knowledge* → Linguistic knowledge for AddSent (Zhou, Huang and Zhu 2020; Wu and Xu 2020), Named Entity Recognition (NER) for NerChanged and RoleSwitched (Mitra *et al*. 2020), Semantic Role Labelling (SRL) for AddSent (Chen and Durrett 2020), Consistency on counterfactual examples for RTE and QA (Teney, Abbasnedjad, and van den Hengel 2020a; Asai and Hajishirzi 2020)
Adversarial Training	*Trains model on data that was generated to maximise the prediction error of the model* → AddSent (Yuan *et al*. 2019; Liu *et al*. 2020a; Liu *et al*. 2020c; Welbl *et al*. 2020); Word Perturbations in RTE (Jia *et al*. 2019); HypOnly (Stacey *et al*. 2020; Liu *et al*. 2020b)
Multi-task learning	*Optimised the model jointly on an additional task that provides additional signal against a weakness* → Explanation Reconstruction for MRC (Rajagopal *et al*. 2020); Paraphrase identification and SRL for HANS (Tu *et al*. 2020; Cengiz and Yuret 2020)

et al. (2020) show that a biased classifier can be approximated by overfitting a regular model on a small portion of the training set. For RTE, Zhang *et al*. (2020a) compare the effects of different proposed debiasing variants discussed in this paragraph. They find that these approaches yield moderate improvements in out-of-distribution performance (up to 7% using the method by He *et al*. 2019).

In an effort to incorporate external knowledge into the model to increase its robustness, multi-task training frameworks with semantic role labelling (SRL) (Cengiz and Yuret 2020) and

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

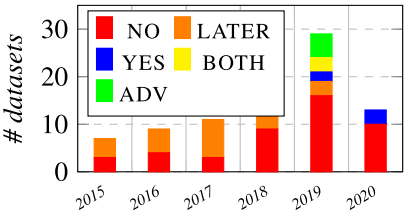

Figure 5. Dataset by publication year with NO or ANY spurious correlations detection methods applied; applied in a LATER publication; created ADVersarially, or BOTH.

explanation reconstruction (Rajagopal *et al.* 2020) have been proposed. It is interesting to note that SRL is a popular choice for incorporating additional linguistic information (Wu *et al.* 2019; Chen and Durrett 2020) due to the fact that it exhibits syntactic and semantic information independent of the specific dataset. Additional external resources encoded into the models during training are named entities (Mitra *et al.* 2020), information from knowledge bases (Wu and Xu 2020) or logic constraints (Minervini and Riedel 2018).

Interestingly, inconsistency on counterexamples, such as those used for training data augmentation, can be explicitly utilised as a regularisation penalty, to encourage models to detect meaningful differences in input data (Teney *et al.* 2020a; Asai and Hajishirzi 2020). Countermeasures for circumventing multi-hop reasoning are providing labels as strong supervision signal for spans that bridge the information between multiple sentences (Jiang and Bansal 2019) or decomposing and sequentially processing compositional questions (Tang *et al.* 2021).

6. Impact on the creation of new datasets

Finally, we report whether the existence of spurious correlations is considered when publishing new resources, by applying any quantitative methods such as those discussed in Section 5.1, or whether some kind of adversarial pruning discussed in Section 5.3 was employed. The results are shown in Figure 5. We observe that the publications we use as our seed papers for the survey (c.f. Section 2) in fact seem to impact how novel datasets are presented, as after their publication (in years 2017 and 2018), a growing number of papers report partial baseline results and existing correlations in their data (four in 2018 and five in 2019). Furthermore, newly proposed resources are increasingly pruned against state-of-the-art approaches (nine in 2018 and 2019 cumulative). However, for nearly half (46 out of 96) of the datasets under investigation, there is no information about potential spurious correlations yet. The scientific community would benefit from an application of the quantitative methods that have been presented in this survey to those NLU datasets.

7. Discussion and conclusion

We present a structured survey of methods that reveal flaws in NLU datasets, methods that show that neural models inherit those correlations or assess their capabilities otherwise, and methods that mitigate those weaknesses. Due to the prevalence of simple, low-bias architectures, the lack of data diversity and existence of data specific artefacts result in models that fail to discriminate between spurious and reliable correlation signals in training data. This, in turn, confounds the hypotheses about the capabilities they acquire when trained and evaluated on these data. More realistic, lower estimates of their capabilities are reported when evaluated on data drawn from a different distribution and with focus on specific capabilities. Efforts towards more robust models include injecting additional knowledge by augmenting training data or introducing constraints into the model architecture, heavier regularisation and training on auxiliary tasks, or encoding more knowledge-intensive input representations.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Based on these insights, we formulate the following recommendations for possible future research directions:

- Most methods discussed in this survey bear negative predictive power only, but the absence of negative results cannot be interpreted as positive evidence. This can be taken as a motivation to put more effort into research that verifies robustness (Shi *et al*. 2020), develops model 'test suites' inspired by good software engineering practices (Ribeiro *et al*. 2020) or provides worst-case performance bounds (Raghunathan, Steinhardt, and Liang 2018; Jia *et al*. 2019). Similar endeavours are pursued by researchers that propose to overthink the empirical risk minimisation (ERM) principle where the assumption is that the performance on the evaluation data can be approximated by the performance on training data, in favour of approaches that relax this assumption. Examples include optimising worst-case performance on a group of training sets (Sagawa *et al*. 2020) or learning features that are invariant in multiple training environments (Teney, Abbasnejad and Hengel 2020b).
- While one of the main themes for combatting reliance on spurious correlations is by injecting additional knowledge, there is a need for a systematic investigation of the type and amount of prior knowledge on neural models' out-of-distribution adversarial and challenge set evaluation performance.
- Partial input baselines are conceptually simple and cheap to employ for any task, so researchers should be encouraged to apply and report their performance when introducing a novel dataset. While not a guarantee for the absence of spurious correlations (Feng et al. 2019), they can hint at their presence and provide more context to quantitative evaluation scores. The same holds true for methods that report existing correlations in data.
- Training set-free, expert-curated evaluation benchmarks that focus on specific phenomena (Linzen 2020) are an obvious way to evaluate capabilities of NLP models without the confounding the effects of spurious correlations between training and test data. Challenge sets discussed in this work, however, measure the performance on the investigated phenomenon on out-of-distribution data and provide informal arguments on why the distribution shift is negligible. How to formally disentangle this effect from the actual capability to process the investigated phenomenon remains an open question.

Specifically for the area of NLU as discussed in this paper, we additionally outline the following recommendations:

- Adapting methods applied to RTE datasets or developing novel methodologies to reveal cues and spurious correlations in MRC data is a possible future research direction.
- The growing number of MRC datasets provides a natural test bed for the evaluation of out-of-distribution generalisation. However, studies concerning this (Talmor and Berant 2019; Fisch *et al*. 2019; Miller *et al*. 2020) mostly focus on empirical experiments. Theoretical contributions, for example, by using the causal inference framework (Magliacane *et al*. 2017), could help to explain their results.
- Additionally, due to its flexibility, the MRC task allows for the formulation of problems that are inherently hard for the state of the art, such as systematic generalisation (Lake and Baroni 2017). Experiments with synthetic data, such as those discussed in this paper, need to be complemented with natural datasets, such as evaluating the understanding of and appropriate reactions to new situations presented in the context. Talmor *et al*. (2020) make a step in this direction.
- While RTE is increasingly becoming a popular task to attribute various reading and reasoning capabilities to neural models, the transfer of those capabilities to different tasks, such as MRC, remains to be seen. Additionally, the MRC task requires further capabilities

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

that cannot be tested in an RTE setting conceptually, such as selecting the relevant answer sentence from distracting context or integrating information from multiple sentences, both shown to be inadequately tested by current state-of-the-art gold standards (Jia and Liang 2017; Jiang and Bansal 2019). Therefore, it is important to develop those challenge sets for MRC models as well in order to gain a more focused understanding of their capabilities and limitations.

It is worth mentioning, that – perhaps unsurprisingly – neural models' notion of complexity does not necessarily correlate with that of humans. In fact, after creating a 'hard' subset of their evaluation data that is clean of spurious correlations, Yu *et al*. (2020) report an increase in human performance, directly contrary to the neural models they evaluate. Partial baseline methods suggest a similar conclusion: without the help of statistics, humans will arguably not be able to infer whether a sentence is entailed by another sentence they never see, whereas neural networks excel at it (Poliak *et al*. 2018; Gururangan *et al*. 2018).

We want to highlight that the availability of multiple large-scale datasets, albeit exhibiting flaws or spurious correlations, together with the methods such as those discussed in this survey are a necessary prerequisite to gain empirically grounded understanding of what the current state-of-the-art NLU models are learning and where they still fail. This gives targeted suggestions when building the next iteration of datasets and model architectures, and therefore advance the research in NLP. While necessary, it remains to be seen whether this iterative process is sufficient to yield systems that are robust enough to perform any given natural language understanding task, the so-called 'general linguistic intelligence' (Yogatama *et al*. 2019).

References

Abacha A.B., **Dinh D. and Mrabet Y.** (2015). Semantic analysis and automatic corpus construction for entailment recognition in medical texts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9105. Springer Verlag, pp. 238–242.

Ahmad W., **Chi J.**, **Tian Y. and Chang K.-W.** (2020). PolicyQA: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 743–749.

Asai A. and Hajishirzi H. (2020). Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 5642–5650.

Bahdanau D., **Cho K.H. and Bengio Y.** (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR.

Bajgar O., **Kadlec R. and Kleindienst J.** (2016). Embracing Data Abundance: BookTest Dataset for Reading Comprehension. arXiv preprint arXiv 1610.00956.

Belinkov Y., **Poliak A.**, **Shieber S.**, **Van Durme B. and Rush A.** (2019). Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 877–891.

Bowman S.R., **Angeli G.**, **Potts C. and Manning C.D.** (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 632–642.

Bras R.L., **Swayamdipta S.**, **Bhagavatula C.**, **Zellers R.**, **Peters M.E.**, **Sabharwal A. and Choi Y.** (2020). Adversarial filters of dataset biases. In Hal Daumé III and Singh A. (eds), *Proceedings of the 37th International Conference on Machine Learning*. PMLR, pp. 1078–1088.

Brown T.B., **Mann B.**, **Ryder N.**, **Subbiah M.**, **Kaplan J.**, **Dhariwal P.**, **Neelakantan A.**, **Shyam P.**, **Sastry G.**, **Askell A.**, **Agarwal S.**, **Herbert-Voss A.**, **Krueger G.**, **Henighan T.**, **Child R.**, **Ramesh A.**, **Ziegler D.M.**, **Wu J.**, **Winter C.**, **Hesse C.**, **Chen M.**, **Sigler E.**, **Litwin M.**, **Gray S.**, **Chess B.**, **Clark J.**, **Berner C.**, **McCandlish S.**, **Radford A.**, **Sutskever I. and Amodei D.** (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv 2005.14165.

Brühlmann F., **Petralito S.**, **Aeschbach L.F. and Opwis K.** (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology* **2**, 100022.

Cai Z., **Tu L. and Gimpel K.** (2017). Pay attention to the ending: Strong neural baselines for the ROC story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 616–622.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Cengiz C. and Yuret D. (2020). Joint training with semantic role labeling for better generalization in natural language inference. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 78–88.

Chen D., Bolton J. and Manning C.D. (2016). A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 4, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2358–2367.

Chen J. and Durrett G. (2019). Understanding dataset design choices for multi-hop reasoning. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4026–4032.

Chen J. and Durrett G. (2020). Robust Question Answering Through Sub-part Alignment. arXiv preprint arXiv 2004.14648.

Chen M., D'Arcy M., Liu A., Fernandez J. and Downey D. (2019). CODAH: An Adversarially-Authored Question Answering Dataset for Common Sense.

Chien T. and Kalita J. (2020). Adversarial analysis of natural language inference systems. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pp. 1–8.

Choi E., He H., Iyyer M., Yatskar M., Yih W.-t., Choi Y., Liang P. and Zettlemoyer L. (2018). QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2174–2184.

Clark C., Lee K., Chang M.-W., Kwiatkowski T., Collins M. and Toutanova K. (2019a). BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2924–2936.

Clark C., Yatskar M. and Zettlemoyer L. (2019b). Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4067–4080.

Clark J.H., Choi E., Collins M., Garrette D., Kwiatkowski T., Nikolaev V. and Palomaki J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics* **8**, 454–470.

Crane M. and Cheriton D.R. (2018). Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics* **6**, 241–252.

Dagan I., Roth D., Sammons M. and Zanzotto F. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies* **6**(4), 1–222.

Dalvi B., Huang L., Tandon N., Yih W.-t. and Clark P. (2018). Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1595–1604.

Demszky D., Guu K. and Liang P. (2018). Transforming Question Answering Datasets Into Natural Language Inference Datasets. arXiv preprint arXiv:1809.02922.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4171–4186.

Dodge J., Ilharco G., Schwartz R., Farhadi A., Hajishirzi H. and Smith N. (2020). Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. arXiv preprint arXiv 2002.0630.

Dua D., Gottumukkala A., Talmor A., Gardner M. and Singh S. (2019a). Comprehensive multi-dataset evaluation of reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 147–153.

Dua D., Wang Y., Dasigi P., Stanovsky G., Singh S. and Gardner M. (2019b). DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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)*, pp. 2368–2378.

Dunn M., Sagun L., Higgins M., Guney V.U., Cirik V. and Cho K. (2017). SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. arXiv preprint arXiv 1704.05179.

Feng S., Wallace E. and Boyd-Graber J. (2019). Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 5533–5538.

Fisch A., Talmor A., Jia R., Seo M., Choi E. and Chen D. (2019). MRQA 2019 Shared Task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1–13.

Gardner M., **Artzi Y.**, **Basmov V.**, **Berant J.**, **Bogin B.**, **Chen S.**, **Dasigi P.**, **Dua D.**, **Elazar Y.**, **Gottumukkala A.**, **Gupta N.**, **Hajishirzi H.**, **Ilharco G.**, **Khashabi D.**, **Lin K.**, **Liu J.**, **Liu N.F.**, **Mulcaire P.**, **Ning Q.**, **Singh S.**, **Smith N.A.**, **Subramanian S.**, **Tsarfaty R.**, **Wallace E.**, **Zhang A. and Zhou B.** (2020). Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1307–1323.

Geiger A., **Cases I.**, **Karttunen L. and Potts C.** (2019). Posing fair generalization tasks for natural language inference. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4484–4494.

Geiger A., **Richardson K. and Potts C.** (2020). Modular Representation Underlies Systematic Generalization in Neural Natural Language Inference Models. arXiv preprint arXiv 2004.14623.

Glockner M., **Shwartz V. and Goldberg Y.** (2018). Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 650–655.

Goodfellow I.J., **Pouget-Abadie J.**, **Mirza M.**, **Xu B.**, **Warde-Farley D.**, **Ozair S.**, **Courville A. and Bengio Y.** (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2672–2680.

Goodwin E., **Sinha K. and O'Donnell T.J.** (2020). Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1958–1969.

Grail Q., **Perez J. and Silander T.** (2018). Adversarial networks for machine reading. *TAL Traitement Automatique des Langues* **59**(2), 77–100.

Gururangan S., **Swayamdipta S.**, **Levy O.**, **Schwartz R.**, **Bowman S. and Smith N.A.** (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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 107–112.

Habernal I., **Wachsmuth H.**, **Gurevych I. and Stein B.** (2018). The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1930–1940.

Han X., **Wallace B.C. and Tsvetkov Y.** (2020). Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. arXiv preprint arXiv 2005.06676.

He H., **Zha S. and Wang H.** (2019). Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 132–142.

Hermann K.M., **Koçiský T.**, **Grefenstette E.**, **Espeholt L.**, **Kay W.**, **Suleyman M. and Blunsom P.** (2015). Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems, pp. 1693–1701.

Holzenberger N., **Blair-Stanek A. and Van Durme B.** (2020). A dataset for statutory reasoning in tax law entailment and question answering. In *NLLP KDD*.

Huang L., **Le Bras R.**, **Bhagavatula C. and Choi Y.** (2019). Cosmos QA: Machine reading comprehension with contextual 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2391–2401.

Jeretic P., **Warstadt A.**, **Bhooshan S. and Williams A.** (2020). Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 8690–8705.

Jia R. and Liang P. (2017). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031.

Jia R., **Raghunathan A.**, **Göksel K. and Liang P.** (2019). Certified robustness to adversarial word substitutions. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4127–4140.

Jiang Y. and Bansal M. (2019). Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2726–2736.

Jin D., **Pan E.**, **Oufattole N.**, **Weng W.-H.**, **Fang H. and Szolovits P.** (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences* **11**(14), 6421.

Jin Q., **Dhingra B.**, **Liu Z.**, **Cohen W. and Lu X.** (2019). PubMedQA: A dataset for biomedical research question answering. 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), Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2567–2577.

Jing Y., Xiong D. and Yan Z. (2019). BiPaR: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2452–2462.

Joshi P., Aditya S., Sathe A. and Choudhury M. (2020). TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 41–55.

Jurczyk T., Zhai M. and Choi J.D. (2016). SelQA: A new benchmark for selection-based question answering. In *Proceedings - 2016 IEEE 28th International Conference on Tools with Artificial Intelligence, ICTAI 2016*, pp. 820–827.

Kamath S., Grau B. and Ma Y. (2018). An adaption of BIOASQ question answering dataset for machine reading systems by manual annotations of answer spans. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 72–78.

Kang D., Khot T., Sabharwal A. and Hovy E. (2018). 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)*, vol. 1, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2418–2428.

Karimi Mahabadi R., Belinkov Y. and Henderson J. (2020). End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 8706–8716.

Kaushik D., Hovy E. and Lipton Z.C. (2020). Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Kaushik D. and Lipton Z.C. (2018). How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 5010–5015.

Kavumba P., Inoue N., Heinzerling B., Singh K., Reisert P. and Inui K. (2019). When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*. Association for Computational Linguistics (ACL), pp. 33–42.

Khashabi D., Khot T. and Sabharwal A. (2020). Natural Perturbation for Robust Question Answering. arXiv preprint arXiv 2004.04849.

Ko M., Lee J., Kim H., Kim G. and Kang J. (2020a). Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1109–1121.

Ko M., Lee J., Kim H., Kim G. and Kang J. (2020b). Look at the First Sentence: Position Bias in Question Answering. arXiv preprint arXiv 2004.14602.

Kwiatkowski T., Palomaki J., Redfield O., Collins M., Parikh A., Alberti C., Epstein D., Polosukhin I., Devlin J., Lee K., Toutanova K., Jones L., Kelcey M., Chang M.-W., Dai A.M., Uszkoreit J., Le Q. and Petrov S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **7**, 453–466.

Lai A. and Hockenmaier J. (2014). Illinois-LH: A denotational and distributional approach to semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 329–334.

Lake B.M. and Baroni M. (2017). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *35th International Conference on Machine Learning, ICML 2018*, vol. 7, pp. 4487–4499.

Lan Z., Chen M., Goodman S., Gimpel K., Sharma P. and Soricut R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Li P., Li W., He Z., Wang X., Cao Y., Zhou J. and Xu W. (2016). Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering. arXiv preprint arXiv 1607.06275.

Liang Y., Li J. and Yin J. (2019). A new multi-choice reading comprehension dataset for curriculum learning. In *Proceedings of Machine Learning Research*, vol. 101. International Machine Learning Society (IMLS), pp. 742–757.

Lin K., Tafjord O., Clark P. and Gardner M. (2019). Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 58–62.

Linzen T. (2020). How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 5210–5217.

Liu K., Liu X., Yang A., Liu J., Su J., Li S. and She Q. (2020a). A robust adversarial training approach to machine reading comprehension. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pp. 8392–8400.

Liu N.F., Schwartz R. and Smith N.A. (2019a). 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:*

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Human Language Technologies, Volume 1 (Long and Short Papers), Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2171–2179.

Liu P., **Du C.**, **Zhao S. and Zhu C.** (2019b). Emotion Action Detection and Emotion Inference: The Task and Dataset. arXiv preprint arXiv 1903.06901.

Liu T., **Zheng X.**, **Chang B. and Sui Z.** (2020b). HypoNLI: Exploring the artificial patterns of hypothesis-only bias in natural language inference. In *Proceedings of The 12th Language Resources and Evaluation Conference*.

Liu X., **Cheng H.**, **He P.**, **Chen W.**, **Wang Y.**, **Poon H. and Gao J.** (2020c). Adversarial Training for Large Neural Language Models. arXiv preprint arXiv 2004.08994.

Longpre S., **Lu Y. and DuBois C.** (2021). On the transferability of minimal prediction preserving inputs in question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1288–1300.

Magliacane S., **van Ommen T.**, **Claassen T.**, **Bongers S.**, **Versteeg P. and Mooij J.M.** (2017). Domain daptation by using causal inference to predict invariant conditional distributions. In Advances in Neural Information Processing Systems, pp. 10846–10856.

Maharana A. and Bansal M. (2020). Adversarial augmentation policy search for domain and cross-lingual generalization in reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 3723–3738.

Mai G., **Janowicz K.**, **He C.**, **Liu S. and Lao N.** (2018). POIReviewQA: A semantically enriched POI retrieval and question answering dataset. In *Proceedings of the 12th Workshop on Geographic Information Retrieval - GIR'18*, pp. 1–2.

McCoy T., **Pavlick E. and Linzen T.** (2019). 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*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 3428–3448.

Mehrabi N., **Morstatter F.**, **Saxena N.**, **Lerman K. and Galstyan A.** (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys* **54**(6), 1–35.

Mihaylov T., **Clark P.**, **Khot T. and Sabharwal A.** (2018). Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2381–2391.

Miller J., **Krauth K.**, **Recht B. and Schmidt L.** (2020). The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6905–6916.

Min J., **McCoy R.T.**, **Das D.**, **Pitler**, **E. and Linzen T.** (2020). Syntactic Data Augmentation Increases Robustness to Inference Heuristics. arXiv preprint arXiv 2004.11999.

Min S., **Wallace E.**, **Singh S.**, **Gardner M.**, **Hajishirzi H. and Zettlemoyer L.** (2019). Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4249–4257.

Min S., **Zhong V.**, **Socher R. and Xiong C.** (2018). Efficient and robust question answering from minimal context over documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1725–1735.

Minervini P. and Riedel S. (2018). Adversarially regularising neural NLI models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 65–74.

Mishra S., **Arunkumar A.**, **Sachdeva B.**, **Bryan**, **C. and Baral C.** (2020). DQI: Measuring Data Quality in NLP. arXiv preprint arXiv 2005.00816.

Mitra A., **Shrivastava I. and Baral C.** (2020). Enhancing natural language inference using new and expanded training data sets and new learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Möller T., **Reina A.**, **Jayakumar R. and Pietsch M.** (2020). COVID-QA: A question answering dataset for COVID-19 | OpenReview. In *ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID)*.

Mostafazadeh N., **Chambers N.**, **He X.**, **Parikh D.**, **Batra D.**, **Vanderwende L.**, **Kohli P. and Allen J.** (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 839–849.

Mu J. and Andreas J. (2020). Compositional explanations of neurons. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*.

Mudrakarta P.K., **Taly A.**, **Sundararajan M. and Dhamdhere K.** (2018). Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1896–1906.

Mullenbach J., **Gordon J.**, **Peng N. and May J.** (2019). Do nuclear submarines have nuclear captains? A challenge dataset for commonsense reasoning over adjectives and objects. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 6051–6057.

Naik A., **Ravichander A.**, **Sadeh N.**, **Rose C. and Neubig G.** (2018). Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 2340–2353.

Nakanishi M., **Kobayashi T. and Hayashi Y.** (2018). Answerable or not: Devising a dataset for extending machine reading comprehension. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 973–983.

Nie Y., **Wang Y. and Bansal M.** (2019). Analyzing compositionality-sensitivity of NLI models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33(01), pp. 6867–6874.

Nie Y., **Williams A.**, **Dinan E.**, **Bansal M.**, **Weston J. and Kiela D.** (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4885–4901.

Ning Q., **Wu H.**, **Han R.**, **Peng N.**, **Gardner M. and Roth D.** (2020). TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1158–1172.

Niven T. and Kao H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4658–4664.

Pampari A., **Raghavan P.**, **Liang J. and Peng J.** (2018). emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2357–2368.

Panenghat M.P., **Suntwal S.**, **Rafique F.**, **Sharp R. and Surdeanu M.** (2020). Towards the necessity for debiasing natural language inference datasets. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, pp. 6883–6888.

Paperno D., **Kruszewski G.**, **Lazaridou A.**, **Pham Q.N.**, **Bernardi R.**, **Pezzelle S.**, **Baroni M.**, **Boleda G.** and **Fernández R.** (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, vol. 3, pp. 1525–1534.

Pappas D., **Stavropoulos P.**, **Androutsopoulos I. and McDonald R.** (2020). BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 140–149.

Pavlick E. and Kwiatkowski T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics* **7**, 677–694.

Poliak A., **Haldar A.**, **Rudinger R.**, **Hu J.E.**, **Pavlick E.**, **White A.S. and Van Durme B.** (2018). Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 67–81.

Pugaliya H., **Route J.**, **Ma K.**, **Geng Y. and Nyberg E.** (2019). Bend but don't break? Multi-challenge stress test for QA models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 125–136.

Raffel C., **Shazeer N.**, **Roberts A.**, **Lee K.**, **Narang S.**, **Matena M.**, **Zhou Y.**, **Li W. and Liu P.J.** (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**, 1–67.

Raghunathan A., **Steinhardt J. and Liang P.** (2018). Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, pp. 10900–10910.

Rajagopal D., **Tandon N.**, **Clark P.**, **Dalvi B. and Hovy E.** (2020). What-if I ask you to explain: Explaining the effects of perturbations in procedural text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 3345–3355.

Rajpurkar P., **Zhang J.**, **Lopyrev K. and Liang P.** (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2383–2392.

Ribeiro M.T., **Guestrin C. and Singh S.** (2019). Are red roses red? Evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 6174–6184.

Ribeiro M.T., **Singh S. and Guestrin C.** (2018). 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)*, vol. 1, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 856–865.

Ribeiro M.T., **Wu T.**, **Guestrin C. and Singh S.** (2020). Beyond accuracy: Behavioral testing of NLP models with checkList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4902–4912.

Richardson K., **Hu H.**, **Moss L.S. and Sabharwal A.** (2019). Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Richardson K. and Sabharwal A. (2020). What does my QA model know? Devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics* **8**, 572–588.

Roemmele M., **Bejan C.A. and Gordon A.S.** (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Rogers A., **Kovaleva O.**, **Downey M. and Rumshisky A.** (2020). Getting closer to AI complete question answering: A set of prerequisite real tasks. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 8722–8731.

Rozen O., **Shwartz V.**, **Aharoni R. and Dagan I.** (2019). Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 196–205.

Rudinger R., **May C. and Van Durme B.** (2017). Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 74–79.

Rychalska B., **Basaj D.**, **Wróblewska A. and Biecek P.** (2018a). Does it care what you asked? Understanding importance of verbs in deep learning QA system. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 322–324.

Rychalska B., **Basaj D.**, **Wróblewska A. and Biecek P.** (2018b). How much should you ask? On the question structure in QA systems. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 319–321.

Sagawa S., **Koh P.W.**, **Hashimoto T.B. and Liang P.** (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*.

Saikh T., **Ekbal A. and Bhattacharyya P.** (2020). ScholarlyRead: A new dataset for scientific article reading comprehension. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, pp. 5498–5504.

Sakaguchi K., **Bras R.L.**, **Bhagavatula C. and Choi Y.** (2020). WinoGrande: An adversarial winograd schema challenge at scale. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 8732–8740.

Salvatore F., **Finger M. and Hirata** Jr **R.** (2019). A logical-based corpus for cross-lingual evaluation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 22–30.

Sanchez I., **Mitchell J. and Riedel S.** (2018). Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1975–1985.

Schick T. and Schütze H. (2021). 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*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2339–2352.

Schlegel V., **Valentino M.**, **Freitas A.A.**, **Nenadic G. and Batista-Navarro R.** (2020). A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 5359–5369.

Schmitt M. and Schütze H. (2019). SherLIiC: A typed event-focused lexical inference benchmark for evaluating natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 902–914.

Schuster T., **Shah D.**, **Yeo Y.J.S.**, **Roberto Filizzola Ortiz D.**, **Santus E. and Barzilay R.** 2019. Towards debiasing fact verification 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 3417–3423.

Schwartz R., **Sap M.**, **Konstas I.**, **Zilles L.**, **Choi Y. and Smith N.A.** (2017). The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 15–25.

Seo M.J., **Kembhavi A.**, **Farhadi A. and Hajishirzi H.** (2017). Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Shi Z., **Zhang H.**, **Chang K.-W.**, **Huang M. and Hsieh C.-J.** (2020). Robustness verification for transformers. In *8th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Si C., **Wang S.**, **Kan M.-Y. and Jiang J.** (2019). What does BERT Learn from Multiple-Choice Reading Comprehension Datasets? arXiv preprint arXiv:1910.12391.

Si C., **Yang Z.**, **Cui Y.**, **Ma W.**, **Liu T. and Wang S.** (2020). Benchmarking Robustness of Machine Reading Comprehension Models. arXiv preprint arXiv 2004.14004.

Sinha K., **Sodhani S.**, **Dong J.**, **Pineau J. and Hamilton W.L.** (2019). CLUTRR: A diagnostic benchmark for inductive reasoning from text. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4505–4514.

Stacey J., **Minervini P.**, **Dubossarsky H.**, **Riedel S. and Rocktäschel T.** (2020). There is Strength in Numbers: Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. arXiv preprint arXiv 2004.07790.

Starc J. and Mladenić D. (2017). Constructing a Natural Language Inference dataset using generative neural networks. *Computer Speech & Language* **46**, 94–112.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Sugawara S., **Inui K.**, **Sekine S. and Aizawa A.** (2018). What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4208–4219.

Sugawara S., **Kido Y.**, **Yokono H. and Aizawa A.** (2017a). Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 806–817.

Sugawara S., **Stenetorp P.**, **Inui K. and Aizawa A.** (2020). Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sugawara S., **Yokono H. and Aizawa A.** (2017b). Prerequisite skills for reading comprehension: Multi-perspective analysis of MCTest datasets and systems. In *Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3089–3096.

Szegedy C., **Zaremba W.**, **Sutskever I.**, **Bruna J.**, **Erhan D.**, **Goodfellow I. and Fergus R.** (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tafjord O., **Clark P.**, **Gardner M.**, **Yih W.-t. and Sabharwal A.** (2018). QuaRel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7063–7071.

Talmor A. and Berant J. (2019). MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4911–4921.

Talmor A., **Tafjord O.**, **Clark P.**, **Goldberg Y. and Berant J.** (2020). Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*.

Tan S., **Shen Y.**, **Huang C.-w. and Courville A.** (2019). Investigating Biases in Textual Entailment Datasets. arXiv preprint arXiv 1906.09635.

Tandon N., **Dalvi B.**, **Sakaguchi K.**, **Clark P. and Bosselut A.** (2019). WIQA: A dataset for "What if. . ." reasoning over procedural text. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 6075–6084.

Tang Y., **Ng H.T. and Tung A.** (2021). Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 3244–3249.

Teney D., **Abbasnedjad E. and van den Hengel A.** (2020a). Learning what makes a difference from counterfactual examples and gradient supervision. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, LNCS, vol. 12355, pp. 580–599.

Teney D., **Abbasnejad E. and van den Hengel A.** (2020b). Unshuffling Data for Improved Generalization. arXiv.

Thorne J., **Vlachos A.**, **Christodoulopoulos C. and Mittal A.** (2018). FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 809–819.

Thorne J., **Vlachos A.**, **Christodoulopoulos C. and Mittal A.** (2019). Evaluating adversarial attacks against multiple fact verification systems. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2944–2953.

Trichelair P., **Emami A.**, **Trischler A.**, **Suleman K. and Cheung J.C.K.** (2019). How reasonable are common-sense reasoning tasks: A case-study on the winograd schema challenge and SWAG. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 3380–3385.

Trivedi H., **Balasubramanian N.**, **Khot T. and Sabharwal A.** (2020). Measuring and Reducing Non-Multifact Reasoning in Multi-hop Question Answering. arXiv preprint arXiv 2005.00789.

Tsuchiya M. (2018). Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Tu L., **Lalwani G.**, **Gella S. and He H.** (2020). An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics* **8**, 621–633.

Utama P.A., **Moosavi N.S. and Gurevych I.** (2020). Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 8717–8729.

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, **Kaiser L. and Polosukhin I.** (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Vilares D. and Gómez-Rodríguez C. (2019). HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 960–966.

Wallace E., **Feng S.**, **Kandpal N.**, **Gardner M. and Singh S.** (2019). 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*

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2153–2162.

Wang A., **Singh A.A.**, **Michael J.**, **Hill F.**, **Levy O. and Bowman S.R.** (2018). Glue: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv preprint arXiv:1804.07461.

Wang Y. and Bansal M. (2018). Robust machine comprehension models via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, vol. 2, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 575–581.

Welbl J., **Minervini P.**, **Bartolo M.**, **Stenetorp P. and Riedel S.** (2020). Undersensitivity in neural reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1152–1165.

Welbl J., **Stenetorp P. and Riedel S.** (2018). Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics* **6**, 287–302.

Wen T.-H., **Vandyke**, **D.**, **Mrkši'c N.**, **Gašić M.**, **Rojas-Barahona L.M.**, **Su P.-H.**, **Ultes S. and Young S.** (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1. Association for Computational Linguistics, pp. 438–449.

Williams A., **Nangia N. and Bowman S.** (2018). 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1112–1122.

Wu B., **Huang H.**, **Wang Z.**, **Feng Q.**, **Yu J. and Wang B.** (2019). Improving the robustness of deep reading comprehension models by leveraging syntax prior. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 53–57.

Wu Z. and Xu H. (2020). Improving the robustness of machine reading comprehension model with hierarchical knowledge and auxiliary unanswerability prediction. *Knowledge-Based Systems* **203**, 106075.

Xiong W., **Wu J.**, **Wang H.**, **Kulkarni V.**, **Yu M.**, **Chang S.**, **Guo X. and Wang W.Y.** (2019). TWEETQA: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 5020–5031.

Xu H., **Ma Y.**, **Liu H.-C.**, **Deb D.**, **Liu H.**, **Tang J. and Jain A.K.** (2020). Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing* **17**(2), 151–178.

Yaghoobzadeh Y., **Tachet R.**, **Hazen T.J. and Sordoni A.** (2019). Robust Natural Language Inference Models with Example Forgetting. arXiv preprint arXiv 1911.03861.

Yanaka H., **Mineshima K.**, **Bekki D.**, **Inui K.**, **Sekine S.**, **Abzianidze L. and Bos J.** (2019a). Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 31–40.

Yanaka H., **Mineshima K.**, **Bekki D.**, **Inui K.**, **Sekine S.**, **Abzianidze L. and Bos J.** (2019b). HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (∗SEM 2019)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 250–255.

Yang Y., **Malaviya C.**, **Fernandez J.**, **Swayamdipta S.**, **Le Bras R.**, **Wang J.-P.**, **Bhagavatula C.**, **Choi Y. and Downey D.** (2020). Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 1008–1025.

Yang Y., **Yih W.-t. and Meek C.** (2015). WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2013–2018.

Yang Z., **Qi P.**, **Zhang S.**, **Bengio Y.**, **Cohen W.W.**, **Salakhutdinov R. and Manning C.D.** (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2369–2380.

Yatskar M. (2019). A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. 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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 2318–2323.

Yogatama D., **D'Autume C.d.M.**, **Connor J.**, **Kocisky T.**, **Chrzanowski**, **M.**, **Kong**, **L.**, **Lazaridou A.**, **Ling W.**, **Yu L.**, **Dyer C. and Blunsom P.** (2019). Learning and Evaluating General Linguistic Intelligence. arXiv preprint arXiv:1901.11373.

Yu W., **Jiang Z.**, **Dong Y. and Feng J.** (2020). ReClor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Yuan F., **Lin Z.**, **Geng Y.**, **Wang W. and Shi G.** (2019). A robust adversarial reinforcement framework for reading comprehension. In *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, pp. 752–759.

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Zellers R., Bisk Y., Schwartz R. and Choi Y. (2018). SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 93–104.

Zellers R., Holtzman A., Bisk Y., Farhadi A. and Choi Y. (2019). HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4791–4800.

Zhang G., Bai B., Liang J., Bai K., Chang S., Yu M., Zhu C. and Zhao T. (2019a). Selection bias explorations and debias methods for natural language sentence matching datasets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 4418–4429.

Zhang G., Bai B., Liang J., Bai K., Zhu C. and Zhao T. (2020a). Reliable Evaluations for Natural Language Inference based on a Unified Cross-dataset Benchmark. arXiv preprint arXiv 2010.07676.

Zhang G., Bai B., Zhang J., Bai K., Zhu C. and Zhao T. (2019b). Mitigating Annotation Artifacts in Natural Language Inference Datasets to Improve Cross-dataset Generalization Ability. arXiv preprint arXiv 1909.04242.

Zhang S., Liu X., Liu J., Gao J., Duh K. and Van Durme B. (2018). ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. arXiv preprint arXiv:1810.12885.

Zhang W.E., Sheng Q.Z., Alhazmi A. and Li C. (2020b). Adversarial attacks on deep-learning models in natural language processing. *ACM Transactions on Intelligent Systems and Technology* **11**(3), 1–41.

Zhou B., Khashabi D., Ning Q. and Roth D. (2019). "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)*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 3361–3367.

Zhou M., Huang M. and Zhu X. (2020). Robust reading comprehension with linguistic constraints via posterior regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**, 2500–2510.

Zhou X. and Bansal M. (2020). Towards robustifying NLI models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 8759–8771.

Appendix A. Inclusion Criteria for the Dataset Corpus

We expand the collection of papers introducing datasets that were investigated or used by any publication in the original survey corpus (e.g., those shown in Figure 1 by a Google Scholar search using the queries shown in Table A1. We include a paper if it introduces a dataset for an NLI task according to our definition and the language of that dataset is English, otherwise we exclude it.

Table A1. Google Scholar Queries for the extended dataset corpus

`allintitle: reasoning ("reading comprehension" OR "machine comprehension") -image -visual -"knowledge graph" -"knowledge graphs"`
`allintitle: comprehension ((((set OR dataset) OR corpus) OR benchmark) OR "gold standard") -image -visual -"knowledge graph" -"knowledge graphs"`
`allintitle: entailment ((((set OR dataset) OR corpus) OR benchmark) OR "gold standard") -image -visual -"knowledge graph" -"knowledge graphs"`
`allintitle: reasoning ((((set OR dataset) OR corpus) OR benchmark) OR "gold standard") -image -visual -"knowledge graph" -"knowledge graphs"`
`allintitle: QA ((((set OR dataset) OR corpus) OR benchmark) OR "gold standard") -image -visual -"knowledge graph" -"knowledge graphs" -"open"`
`allintitle: NLI ((((set OR dataset) OR corpus) OR benchmark) OR "gold standard") -image -visual -"knowledge graph" -"knowledge graphs"`
`allintitle: language inference ((((set OR dataset) OR corpus) OR benchmark) OR "gold standard") -image -visual -"knowledge graph" -"knowledge graphs"`
`allintitle: "question answering" ((((set OR dataset) OR corpus) OR benchmark) OR "gold standard") -image -visual -"knowledge graph" -"knowledge graphs"`

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

Table B1. Table of datasets where no quantitative methods that describe dataset weaknesses have been applied yet

Year	Dataset
2015	MedlineRTE (Abacha, Dinh, and Mrabet 2015), WikiQA (Yang, Yih, and Meek 2015), DailyMail (Hermann *et al.* 2015)
2016	WebQA (Li *et al.* 2016), BookTest (Bajgar, Kadlec, and Kleindienst 2016), SelQA (Jurczyk, Zhai, and Choi 2016), LAMBADA (Paperno *et al.* 2016)
2017	SearchQA (Dunn *et al.* 2017), GANNLI (Starc and Mladenić 2017), CambridgeDialogs (Wen *et al.* 2017)
2018	emrQA (Pampari *et al.* 2018), PoiReviewQA (Mai *et al.* 2018), ReCoRd (Zhang *et al.* 2018), QuAC (Choi *et al.* 2018), ProPara (Dalvi *et al.* 2018), BioASQ (Kamath, Grau, and Ma 2018), OBQA (Mihaylov *et al.* 2018), MedHop (Welbl, Stenetorp, and Riedel 2018), QuaRel (Tafjord *et al.* 2018)
2019	WIQA (Tandon *et al.* 2019), BiPaR (Jing, Xiong, and Yan 2019), PubMedQA (Jin *et al.* 2019), ROPES (Lin *et al.* 2019), HELP (Yanaka *et al.* 2019b), MCTACO (Zhou *et al.* 2019), TWEET-QA (Xiong *et al.* 2019), CosmosQA (Huang *et al.* 2019), RACE-C (Liang, Li, and Yin 2019), SherLIiC (Schmitt and Schütze 2019), VGNLI (Mullenbach *et al.* 2019), NaturalQ (Kwiatkowski *et al.* 2019), BoolQ (Clark *et al.* 2019a), CEAC (Liu *et al.* 2019b), HEAD-QA (Vilares and Gómez-Rodríguez 2019), CLUTRR (Sinha *et al.* 2019)
2020	QuAIL (Rogers *et al.* 2020), BioMRC (Pappas *et al.* 2020), ScholarlyRead (Saikh, Ekbal, and Bhattacharyya 2020), COVID-QA (Möller *et al.* 2020), SARA (Holzenberger, Blair-Stanek, and Van Durme 2020), TORQUE (Ning *et al.* 2020), MedQA (Jin *et al.* 2021), MKQA (Longpre *et al.* 2021), TyDiQA (Clark *et al.* 2020), PolicyQA (Ahmad *et al.* 2020)

Appendix B. Detailed Survey Results

Table B1 shows those 45 datasets from Figure 5 broken down by year, where no quantitative methods to describe possible spurious correlations have been applied yet.

The following table shows the full list of surveyed papers, grouped by dataset and method applied. As papers potentially report the application of multiple methods on multiple datasets, they can appear in the table more than once:

Dataset	Method used	Used by/Investigated by
MNLI	Adversarial Evaluation	Han *et al.* (2020), Chien and Kalita (2020), Nie *et al.* (2019)
	Challenge set	Rozen *et al.* (2019), Liu *et al.* (2019b), Glockner *et al.* (2018), Richardson *et al.* (2019), Nie *et al.* (2019), McCoy *et al.* (2019), Naik *et al.* (2018)
	Arch/Training Improvements	Sagawa *et al.* (2020), Stacey *et al.* (2020), Minervini and Riedel (2018), He *et al.* (2019), Mitra *et al.* (2020), Yaghoobzadeh *et al.* (2019), Wang *et al.* (2018), Zhou and Bansal (2020), Clark et al. (2019b), Zhang *et al.* (2020b), Karimi Mahabadi *et al.* (2020), Belinkov *et al.* (2019)
	Heuristics	Gururangan *et al.* (2018), Tan *et al.* (2019), Poliak *et al.* (2018), Zhang *et al.* (2019a), Bras *et al.* (2020), Nie *et al.* (2019), McCoy *et al.* (2019)
	Partial Baselines	Gururangan *et al.* (2018), Poliak *et al.* (2018), Nie *et al.* (2019)
	Data Improvements	Mitra *et al.* (2020), Panenghat *et al.* (2020), Zhou and Bansal (2020), Min *et al.* (2020)

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

	Manual Analyses	Pavlick and Kwiatkowski (2019)
SQuAD	Adversarial Evaluation	Jia and Liang (2017), Rychalska *et al*. (2018b), Mudrakarta *et al*. (2018), Rychalska *et al*. (2018a), Wallace *et al*. (2019)
	Arch/Training Improvements	Min *et al*. (2018), Zhou *et al*. (2019), Yuan *et al*. (2019), Liu *et al*. (2020a), Wu and Xu (2020), Ko *et al*. (2020b), Clark et al. (2019b), Wu *et al*. (2019)
	Manual Analyses	Sugawara *et al*. (2017a), Pugaliya *et al*. (2019), Sugawara *et al*. (2018)
	Heuristics	Ko *et al*. (2020b), Sugawara *et al*. (2018)
	Data Improvements	Wang and Bansal (2018), Nakanishi *et al*. (2018)
	Partial Baselines	Kaushik and Lipton (2018), Sugawara *et al*. (2020)
	Challenge set	Liu *et al*. (2019b), Ribeiro *et al*. (2019), Nakanishi *et al*. (2018), Dua *et al*. (2019a)
FEVER	Adversarial Evaluation	Thorne *et al*. (2019)
	Data Improvements	Panenghat *et al*. (2020), Schuster *et al*. (2019)
	Heuristics	Schuster *et al*. (2019)
	Arch/Training Improvements	Schuster *et al*. (2019)
ARCT	Heuristics	Niven and Kao (2019)
	Adversarial Evaluation	Niven and Kao (2019)
SWAG	Data Improvements	Zellers *et al*. (2018), Zellers *et al*. (2019)
	Partial Baselines	Trichelair *et al*. (2019), Sugawara *et al*. (2020)
RACE	Adversarial Evaluation	Si *et al*. (2019, 2020)
	Partial Baselines	Si *et al*. (2019), Sugawara *et al*. (2020)
	Heuristics	Sugawara *et al*. (2018)
	Manual Analyses	Sugawara *et al*. (2018)
DREAM	Partial Baselines	Si *et al*. (2019)
	Adversarial Evaluation	Si *et al*. (2019)
MCScript	Partial Baselines	Si *et al*. (2019)
	Adversarial Evaluation	Si *et al*. (2019)
	Heuristics	Sugawara *et al*. (2018)
	Manual Analyses	Sugawara *et al*. (2018)
MCScript 2.0	Partial Baselines	Si *et al*. (2019)
	Adversarial Evaluation	Si *et al*. (2019)
MCTest	Partial Baselines	Si *et al*. (2019), Sugawara *et al*. (2020)
	Adversarial Evaluation	Si *et al*. (2019)
	Manual Analyses	Sugawara *et al*. (2017a,b); Sugawara *et al*. (2018)
	Heuristics	Sugawara *et al*. (2018)

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

DROP	Data Improvements	Dua *et al*. (2019b)
	Manual Analyses	Schlegel *et al*. (2020)
	Challenge set	Gardner *et al*. (2020), Dua *et al*. (2019a)
ANLI	Data Improvements	Nie *et al*. (2020)
HellaSWAG	Data Improvements	Zellers *et al*. (2019)
SNLI	Adversarial Evaluation	Sanchez *et al*. (2018), Nie *et al*. (2019)
	Heuristics	Rudinger *et al*. (2017), Mishra *et al*. (2020), Gururangan *et al*. (2018), Tan *et al*. (2019), Poliak *et al*. (2018), Zhang *et al*. (2019a), Bras *et al*. (2020), Nie *et al*. (2019)
	Arch/Training Improvements	Stacey *et al*. (2020), Minervini and Riedel (2018), Jia *et al*. (2019), He *et al*. (2019), Mitra *et al*. (2020), Zhang *et al*. (2020b), Karimi Mahabadi *et al*. (2020), Belinkov *et al*. (2019)
	Data Improvements	Mishra *et al*. (2020), Mitra *et al*. (2020), Kang *et al*. (2018), Kaushik *et al*. (2020)
	Partial Baselines	Gururangan *et al*. (2018), Poliak *et al*. (2018), Feng et al. (2019), Nie *et al*. (2019), Tsuchiya (2018)
	Manual Analyses	Pavlick and Kwiatkowski (2019)
	Challenge set	Glockner *et al*. (2018), Richardson *et al*. (2019), Nie *et al*. (2019), Kaushik *et al*. (2020)
HotPotQA	Adversarial Evaluation	Jiang and Bansal (2019)
	Data Improvements	Jiang and Bansal (2019)
	Arch/Training Improvements	Jiang and Bansal (2019)
	Manual Analyses	Schlegel *et al*. (2020), Pugaliya *et al*. (2019)
	Partial Baselines	Min *et al*. (2019), Sugawara *et al*. (2020), Chen and Durrett (2019), Trivedi *et al*. (2020)
	Challenge set	Trivedi *et al*. (2020)
	Heuristics	Trivedi *et al*. (2020)
NewsQA	Arch/Training Improvements	Min *et al*. (2018)
	Manual Analyses	Schlegel *et al*. (2020), Sugawara *et al*. (2017a); Sugawara *et al*. (2018)
	Challenge set	Dua *et al*. (2019a)
	Heuristics	Sugawara *et al*. (2018)
TriviaQA	Arch/Training Improvements	Min *et al*. (2018), Clark et al. (2019b)
	Heuristics	Sugawara *et al*. (2018)
	Manual Analyses	Sugawara *et al*. (2018)
HELP	Data Improvements	Yanaka *et al*. (2019b)
ADD-1	Arch/Training Improvements	Stacey *et al*. (2020), Belinkov *et al*. (2019)

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

	Heuristics	Poliak *et al*. (2018)
	Partial Baselines	Poliak *et al*. (2018)
DPR	Arch/Training Improvements	Stacey *et al*. (2020), Belinkov *et al*. (2019)
	Heuristics	Poliak *et al*. (2018)
	Partial Baselines	Poliak *et al*. (2018)
FN+	Arch/Training Improvements	Stacey *et al*. (2020), Belinkov *et al*. (2019)
	Heuristics	Poliak *et al*. (2018)
	Partial Baselines	Poliak *et al*. (2018)
JOCI	Arch/Training Improvements	Stacey *et al*. (2020), Zhang *et al*. (2020b), Belinkov *et al*. (2019)
	Heuristics	Poliak *et al*. (2018)
	Partial Baselines	Poliak *et al*. (2018)
	Manual Analyses	Pavlick and Kwiatkowski (2019)
MPE	Arch/Training Improvements	Stacey *et al*. (2020), Belinkov *et al*. (2019)
	Heuristics	Poliak *et al*. (2018)
	Partial Baselines	Poliak *et al*. (2018)
SICK	Arch/Training Improvements	Stacey *et al*. (2020), Wang *et al*. (2018), Zhang *et al*. (2020b), Belinkov *et al*. (2019)
	Heuristics	Poliak *et al*. (2018), Zhang *et al*. (2019a)
	Partial Baselines	Poliak *et al*. (2018), Lai and Hockenmaier (2014), Tsuchiya (2018)
SPR	Arch/Training Improvements	Stacey *et al*. (2020), Belinkov *et al*. (2019)
	Heuristics	Poliak *et al*. (2018)
	Partial Baselines	Poliak *et al*. (2018)
SciTail	Arch/Training Improvements	Stacey *et al*. (2020), Belinkov *et al*. (2019)
	Heuristics	Poliak *et al*. (2018)
	Partial Baselines	Poliak *et al*. (2018)
	Challenge set	Glockner *et al*. (2018)
MSMarco	Manual Analyses	Schlegel *et al*. (2020), Sugawara *et al*. (2017a), Pugaliya *et al*. (2019), Sugawara *et al*. (2018)
	Heuristics	Sugawara *et al*. (2018)
MultiRC	Manual Analyses	Schlegel *et al*. (2020)
	Partial Baselines	Sugawara *et al*. (2020)
ReCoRd	Manual Analyses	Schlegel *et al*. (2020)

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

COPA	Heuristics	Kavumba *et al*. (2019)
	Challenge set	Kavumba *et al*. (2019)
	Adversarial Evaluation	Kavumba *et al*. (2019)
ReClor	Heuristics	Yu *et al*. (2020)
QA4MRE	Manual Analyses	Sugawara *et al*. (2017a)
Who-did-What	Manual Analyses	Sugawara *et al*. (2017a)
	Partial Baselines	Kaushik and Lipton (2018)
DNC	Manual Analyses	Pavlick and Kwiatkowski (2019)
RTE2	Manual Analyses	Pavlick and Kwiatkowski (2019)
CBT	Arch/Training Improvements	Grail, Perez, and Silander (2018)
	Partial Baselines	Kaushik and Lipton (2018)
CambridgeDialogs	Arch/Training Improvements	Grail *et al*. (2018)
CNN	Partial Baselines	Kaushik and Lipton (2018)
	Manual Analyses	Chen *et al*. (2016)
bAbI	Partial Baselines	Kaushik and Lipton (2018)
ROCStories	Partial Baselines	Schwartz *et al*. (2017), Cai *et al*. (2017)
	Heuristics	Cai *et al*. (2017)
DailyMail	Manual Analyses	Chen *et al*. (2016)
SearchQA	Manual Analyses	Pugaliya *et al*. (2019)
QNLI	Heuristics	Bras *et al*. (2020)
CoQA	Manual Analyses	Yatskar (2019)
	Partial Baselines	Sugawara *et al*. (2020)
QuAC	Manual Analyses	Yatskar (2019)
SQuAD 2.0	Manual Analyses	Yatskar (2019)
	Partial Baselines	Sugawara *et al*. (2020)
	Challenge set	Dua *et al*. (2019a)
DuoRC	Partial Baselines	Sugawara *et al*. (2020)
	Challenge set	Dua *et al*. (2019a)
WikiHop	Partial Baselines	Chen and Durrett (2019)
	Heuristics	Sugawara *et al*. (2018)
	Manual Analyses	Sugawara *et al*. (2018)
ARC	Challenge set	Richardson and Sabharwal (2020)
	Heuristics	Sugawara *et al*. (2018)
	Manual Analyses	Sugawara *et al*. (2018)

https://doi.org/10.1017/S1351324922000171 Published online by Cambridge University Press

OBQA	Challenge set	Richardson and Sabharwal (2020)
BoolQ	Challenge set	Gardner *et al*. (2020)
MCTACO	Challenge set	Gardner *et al*. (2020)
Quoref	Challenge set	Gardner *et al*. (2020), Dua *et al*. (2019a)
ROPES	Challenge set	Gardner *et al*. (2020), Dua *et al*. (2019a)
NarrativeQA	Challenge set	Dua *et al*. (2019a)
	Heuristics	Sugawara *et al*. (2018)
	Manual Analyses	Sugawara *et al*. (2018)
None		Geiger *et al*. (2019), Yanaka *et al*. (2019a), Ribeiro *et al*. (2018), Goodwin *et al*. (2020), Salvatore *et al*. (2019).

Cite this article: Schlegel V, Nenadic G and Batista-Navarro R (2023). A survey of methods for revealing and overcoming weaknesses of data-driven Natural Language Understanding. *Natural Language Engineering* **29**, 1–31. https://doi.org/10.1017/S1351324922000171