Beyond Leaderboards: A survey of methods for revealing and overcoming weaknesses in Natural Language Inference data and models

Anonymous COLING submission

Abstract

Recent years have seen a growing number of publications that analyse Natural Language Inference (NLI) 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 it will be a useful resource for researchers who propose new datasets, to have a set of tools to assess the suitability and quality of their data to evaluate various phenomena of interest, as well as those who develop novel architectures, to further understand the implications of their improvements with respect to their model's acquired capabilities.

1 Introduction

Research in areas that require natural language inference (NLI) over text, such as Recognizing Textual Entailment (RTE) (Dagan et al., 2006) and Machine Reading Comprehension (MRC) is advancing at an unprecedented rate. On the one hand, novel architectures (Vaswani et al., 2017) enable efficient unsupervised training on large corpora to obtain expressive contextualised word and sentence representations for a multitude of downstream NLP tasks (Devlin et al., 2019). On the other hand, large-scale datasets (Bowman et al., 2015; Rajpurkar et al., 2016; Williams et al., 2018) provide sufficient examples to optimise large neural models that are capable of outperforming the human baseline on multiple tasks (Raffel et al., 2019; Lan et al., 2020).

Recent work, however, has questioned the seemingly superb performance for some of the tasks. Specifically, training and evaluation data may contain exploitable superficial cues, such as syntactic patterns (McCoy et al., 2019), specific words (Poliak et al., 2018) or sentence length (Gururangan et al., 2018) that are predictive of the expected output. After having been evaluated on data in which those cues have been removed, the performance of those models deteriorated significantly (McCoy et al., 2019; Niven and Kao, 2019), showing that they are in fact relying on the existing cues rather than learning to understand meaning or perform inference. In other words, those well-performing models tend to obtain optimal performance on a particular dataset, i.e. overfitting on it, rather than generalising for the underlying task. This issue, in fact, remains concealed, if a model is compared to a human baseline by means of a single number that reports the average score on a held-out test set, which is typically the case with contemporary benchmark leaderboards.

To reveal and overcome these issues mentioned above, a growing number of approaches has been proposed in the past. All those methods contribute towards a fine-grained understanding of whether the existing methodology actually evaluates the required inference capabilities, what existing models learn from available training data and, more importantly, which capabilities they still fail to acquire, thus providing targeted suggestions for future research.

To make sense of this growing body of literature and help researchers new to the field to navigate it, we present a structured survey of the recently proposed methods and report the trends, applications

Heuristic	E	$\neg E$
Lex. Overlap	2,158	261
Subsequence	1,274	72
Constituent	1,004	58

Figure 1: Number of premise-hypothesis pairs in an RTE dataset following lexical patterns, spuriously skewed towards *Entailment* (McCoy et al., 2019).

Paragraph: *"[. . .] The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV."*
Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*
Original model prediction: John Elway
Model prediction after inserting a distracting sentence: Jeff Dean

Figure 2: Models' over-stability towards common words in question and paragraph, revealed by adversarially inserting distracting sentences (Jia and Liang, 2017).

and findings. In the remainder of this paper, we first establish terminology, set the objectives and the scope of the survey and describe the data collection methodology. We then present a categorisation of the surveyed methods with their main findings, and finally discuss the arising trends and open research questions.

1.1 Terminology

Tasks: The task of *Recognising Textual Entailment (RTE)* is to decide, for a pair of natural language sentences (premise and hypothesis), whether given the premise the hypothesis is true (*Entailment*), false (*Contradiction*) or whether the two sentences are unrelated (*Neutral*) (Dagan et al., 2013).

We refer to the task of finding the correct answer to a question over a passage of text as *Machine Reading Comprehension (MRC)*, also known as Question Answering (QA). Usual formulations of the task require to predict a span from the passage, select from a given set of alternatives or generate a free-form string (Liu et al., 2019b).

In this paper, we use the term "NLI" in its broader sense, referring to the requirement to perform inference over natural language. Thus we expand the usual textual entailment-based definition to also include MRC, as answering a question can be framed as finding an answer that is entailed by the question and the provided context, and the tasks can be transformed vice versa (Demszky et al., 2018).

Spurious Correlations: We call correlations between input data and the expected prediction as "spurious" if they are not indicative for the underlying task but rather an artefact of the data at hand (as illustrated in Figure 1). They also referred to as "(annotation) artefacts" or "(dataset) biases" in literature. The exploitation of those correlations in order to produce the expected prediction is known as the "Clever Hans Effect", named after a horse that was believed to perform arithmetic tasks but was shown to react to subtle body language cues of the asking person (Pfungst and Rahn, 1911).

Model and Architecture: We refer to the neural network architecture of a model as "architecture", e.g. BiDAF (Seo et al., 2017). We refer to a (statistical) model of a certain architecture that was optimised on a given set of training data simply as "model". It is important to make this distinction, as an optimised model's systematic failures can either be explained by considering the training data (and can potentially be different for a model optimised on different data) or affiliated with the model class (and exist for all models with the same architecture) (Liu et al., 2019a; Geiger et al., 2019).

Stress-test: The evaluation of trained models and neural architectures in a controlled way with regard to a phenomenon of interest, such as a reasoning type (e.g. logic inference (Richardson and Sabharwal, 2019)) or a linguistic capability (e.g. lexical semantics (Naik et al., 2018)) is referred to as "stress-testing" (Naik et al., 2018). Measuring the prediction performance of a model with a particular architecture that was trained on a particular dataset on an evaluation-only stress-test (Glockner et al., 2018) allows to draw conclusions about the capabilities the model obtains from the training data. Stress-tests with a training set allow for more general conclusions whether a model with a specific architecture is capable of obtaining the capability, even when optimised with sufficient examples (Kaushik et al., 2020; Geiger et al., 2019).

Out of Distribution: While, traditionally, evaluation of machine learning algorithms is performed under the assumption that training and evaluation data stem from the same generative process and thus are independent and identically distributed, or i.i.d., out of distribution evaluation relaxes this assumption. In NLP, the generative process is determined by the data collection method, usually crowd-sourcing (Rajpurkar et al., 2016; Williams et al., 2018), therefore out of distribution evaluation refers to evaluating on data that are altered in a systematic way (Si et al., 2020) or collected with a different method, e,g. from a different dataset (Dua et al., 2019a; Glockner et al., 2018).

Adversarial: 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 a class of models, while not impacting the human baseline. Examples include appending irrelevant information (Jia and Liang, 2017), illustrated in Figure 2, or paraphrasing (Ribeiro et al., 2019).

Robustness: 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 (unknown) distribution of the evaluation data when compared to the training data, such as introduced by adversarial evaluation or stress-tests, if the out-of-distribution performance of the model is similar to that on the original evaluation set. The opposite of robustness is referred to as "brittleness".

1.2 Objectives and Scope

We aim to provide a comprehensive overview of issues in NLI data and models that are trained and evaluated upon them as well as the methodology used to report them. We set out to address the following questions:
(1) Which NLI tasks and corresponding datasets have been investigated?
(2) Which types of weaknesses have been reported in NLI 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 of novel datasets (that were described in published papers)?

1.3 Data collection 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 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 and works that have cited it according to Google Scholar and include papers that describe methods and/or their applications to report either (1) qualitative evaluation of training and/or test data; (2) superficial cues present in data and the tendency of models to pick them up; (3) systematic issues with task formulations and/or data collection methods; (4) analysis of specific linguistic and reasoning phenomena in data and/or models' performance on them; or (5) enhancements of models' architecture or training procedure in order to overcome data-specific or model-specific issues, related to phenomena and cues described above. We exclude a paper if its target task does not fall under the NLI definition established above, was published before the year 2014 or the language of the target dataset is not English; otherwise, we add it to the set of surveyed papers. With this approach we obtain a total of 87 papers (as of 17^{th} June 2020) from the years 2014-2017 (8), 2018 (18), 2019 (39) and 2020 (22). More than two thirds (53) of the papers were published in venues hosted by the the Association for Computational Linguistics, whereas seven and three were presented in AAAI and ICLR conferences, respectively. The remaining papers were published in other venues (4) or are available as an arXiv preprint (20). 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 final question, we took 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 NLI dataset resource papers (for detailed inclusion and exclusion criteria, see Appendix B). This approach yielded 81 papers. 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 discuss their fairness. For an overview thereof, we refer the interested reader to respective surveys conducted by Zhang et al. (2019c) or Xu et al. (2019) for the first, and by Mehrabi et al. (2019) for the latter.

2 Weaknesses in NLI data and models

Here, we report the types of weaknesses that have found in state-of-the-art NLI data and models.

2.1 Data

We identify three main types of weakness found in the data that was utilised in training and evaluating models and outline them below:

Spurious Correlations In span extraction tasks such as MRC, question (Rychalska et al., 2018), passage wording and the position of the answer span in the passage is indicative of the expected answer for various datasets (Kaushik and Lipton, 2018). In the ROC stories dataset, (Mostafazadeh et al., 2016) where the task is to choose the most plausible ending to a story, the endings exhibit exploitable cues (Schwartz et al., 2017). These cues are even noticeable by humans (Cai et al., 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 and its length correlate with the expected label for a multitude of RTE datasets. The latter study referred to these correlations as "annotation artifacts". 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. Beyond RTE, the choices in the COPA (Roemmele et al., 2011) dataset, where the task is to finish a given passage (similar to ROC Stories), and ARCT (Habernal et al., 2018) where the task is to select whether a statement warrants a claim, contain words that correlate with the expected prediction (Kavumba et al., 2019; Niven and Kao, 2019).

Task unsuitability Chen and Durrett (2019a) demonstrated that selecting from answers in a multiple choice setting considerably simplifies the task when compared to selecting a span from the context. They further showed that for large parts of the popular HotpotQA dataset the answer can be found when deliberately not integrating information from multiple sentences ("multi-hop" reasoning), replicated by Min et al. (2019).

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 or the fact that its correctness is debateable. Neural networks are, in fact, unable to pick up such uncertainty. Furthermore, both Schlegel et al. (2020) and Pugaliya et al. (2019) report the existence of factual errors in MRC evaluation data, where the expected answer to a question is actually wrong. Finally, Rudinger et al. (2017) show the presence of gender and racial stereotypes in crowd-sourced RTE datasets.

2.2 Models

These data weaknesses contribute to brittleness in trained models themselves. Below, we outline those and other issues reported in the literature:

Inheritance of data-related issues Given the issues reported in NLI data, such as spurious correlations, it is worthwhile knowing whether models optimised on this data actually inherit them. In fact, multiple studies confirm this hypothesis, demonstrating that evaluating models on a version of the same dataset where the correlations do not exist, results in poor prediction performance (McCoy et al., 2019; Niven and Kao, 2019; Kavumba et al., 2019).

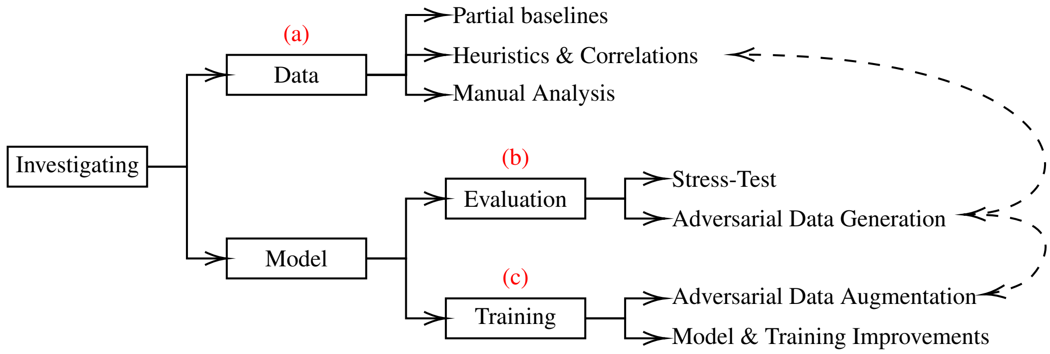

Figure 3: Taxonomy of investigated methods. Dashed arrows indicate conceptually related types of methods, i.e. a method of one type are commonly applied with another method of the related type. Labels (a), (b) and (c) correspond to the coarse grouping discussed in Section 3.

Semantic Over-stability Another weakness, particularly shown for MRC models, is that they appear to not capture the semantics of text beyond superficial lexical features. Neural models struggle to distinguish important from irrelevant sentences that share words with the question (Jia and Liang, 2017), disregard syntactic structure (Basaj et al., 2018; Rychalska et al., 2018) and important words (Mudrakarta et al., 2018) and give inconsistent answer to semantically equivalent questions (Ribeiro et al., 2019). For RTE, they may disregard the composition of the sentence pairs (Nie et al., 2019a).

Generalisation Issues These issues outlined above hint at limited generalisation capabilities of state-of-the-art models that remain concealed using the traditional machine learning evaluation methodology under the in-distribution assumption. Poor evaluation performance on out-of-distribution data (Glockner et al., 2018; Naik et al., 2018; Yanaka et al., 2019b) suggests that current (mostly transformer (Vaswani et al., 2017) based) models are "lazy learners": when possible, they infer simple decision strategies from training data (e.g. based on spurious correlations) that are not representative of the corresponding task instead of learning the required capabilities to perform inference.

3 Methods that reveal and overcome weaknesses in NLI

In the following section we categorise the surveyed papers, briefly describe the categories and illustrate the methodologies by reference 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 above, (b) investigate what inference and reasoning capabilities models optimised on these data acquire when evaluated on samples not drawn from the training distribution and (c) propose architectural (Sagawa et al., 2020) and training procedure (Wang and Bansal, 2018) improvements in order to achieve more robust generalisation beyond data drawn from the training distribution. A schematic overview of the taxonomy of the categories is shown in Figure 3.

3.1 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 NLI 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 gain a deeper understanding of the assessed capabilities (Chen et al., 2016) fall under this category as well. In the analysed body of work, we identified the following three types of methods:

Partial Baselines These methods seek to verify that every input modality 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). Training and evaluating a classifier on parts of the input only suggests that those parts exhibit cues that correlate with the expected prediction, if the measured performance is significantly higher than randomly guessing. Both Gururangan et al. (2018) and Poliak et al. (2018) demonstrated near state-of-the-art performance on multiple RTE datasets, such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), when training a classifier with hypothesis-only input. Kaushik and Lipton (2018) even surpass state-of-the-art MRC models on various datasets when training and evaluating only on parts of the provided input. Methods that mask, drop or shuffle input words or sentences fall under this category as well. Using them, Sugawara et al. (2020) reach performance comparable to that of a model that is trained on full input on a variety of MRC datasets. Similarly, Nie et al. (2019a) reach near state-of-the-art performance on the SNLI and MNLI datasets when shuffling the words in the premise and hypothesis.

Finally, we include methods here that seek to verify whether the data or task formulation is fit to evaluate a particular capability, as they involve training models that are architecturally restricted to obtain said capability, e.g. models that process documents strictly independently to answer questions that require information synthesis from multiple documents (Min et al., 2019; Chen and Durrett, 2019a). Good performance of those impaired models indicates that the task can be solved without the required capability to a certain extent.

Above-chance performance of partial input baselines hints at spurious correlations in the data and suggests that models learn to exploit them; it does not however reveal their precise nature. The opposite does not hold true either: near-chance performance on partial input does not warrant cue-free data, as Feng et al. (2019) illustrate on synthetic examples and published datasets.

Heuristics and Correlations These aim to unveil specific cues and spurious correlations between input and expected output that enable models to learn the task more easily. For sentence pair classification tasks, Gururangan et al. (2018) use the PMI measure between words in a hypothesis and the expected label, while Poliak et al. (2018) use the conditional probability of a label given a word. In contrast, Tan et al. (2019) use word bigrams instead of single words to model their correlation. McCoy et al. (2019) count instances of (subsequently) overlapping words and mutual subtrees of the syntactic parses in a given premise and hypothesis pair, and show that their label distribution is heavily skewed towards entailment. Nie et al. (2019a) optimise a logistic regression model on lexical features and use its confidence to predict a wrong label for a given premise-hypothesis pair as a score for the requirement of inference beyond lexical matching. Niven and Kao (2019) define *productivity* and *coverage* to measure how likely and for what proportion of the dataset an *n*-gram is indicative of the expected label. Cai et al. (2017) propose simple rules based on length, negation and off-the-shelf sentiment analyser scores to select the most probable ending for the ROC story completion task. To show that models actually learn to react to the cues, the data analysis is usually followed by an evaluation on an evaluation set where those correlations are not present anymore (c.f. Section 3.2: Adversarial Evaluation).

Manual Analyses These methods intend to qualitatively analyse the data, if automated approaches as those mentioned above are unsuitable due to the complexity of the phenomena of interest. By means of manual annotation, multiple MRC datasets are reported to lack certain linguistic features and required reasoning and comprehension skills (Schlegel et al., 2020; Sugawara et al., 2017a) which in turn potentially undermines the complexity of the task (Sugawara et al., 2018).

3.2 Model-investigating Methods

Rather than analysing data, approaches described in this section directly evaluate the models with respect to their inference capabilities with regard to various phenomena of interest.

Stress-test is an increasingly popular way to assess trained models and architectures. Naik et al. (2018) automatically generate NLI evaluation data based on an analysis of observed state-of-the-art model error patterns, introducing the term "stress-test". Stress-tests have since been proposed to evaluate the capabilities of handling monotonicity (Yanaka et al., 2019a), lexical inference (Glockner et al., 2018), definitions (Richardson and Sabharwal, 2019) and compositionality (Nie et al., 2019a) for RTE models and semantic equivalence (Ribeiro et al., 2019) for MRC. Liu et al. (2019a) propose to fine-tune a stress-tested model on portions of the stress-test data and observe the change in evaluation performance: if it increases, then the training data did not feature the investigated phenomenon, in the opposite case, however, the model's architecture is unsuited to learn to process the phenomenon alltogther.

Adversarial Evaluation introduces evaluation data that was generated with the aim to "fool" models. Jia and Liang (2017) append question paraphrases to passages of the SQuAD dataset that do not alter the semantics (thus preserving the expected answer) in order to show model over-sensitivity to certain keywords. Methods that evaluate whether models that are trained on data exhibiting spurious correlations learn to exploit those those belong to this category as well. This is usually done by creating a balanced evaluation set, where the spurious correlations present in training data do not exist anymore (McCoy et al., 2019; Kavumba et al., 2019; Niven and Kao, 2019). We further include methods in this category that employ adversarial techniques to interpret a specific model behaviour (Ribeiro et al., 2018; Sanchez et al., 2018; Han et al., 2020). Among those we highlight the work by Wallace et al. (2019), who showed that adversaries generated against a target model tend to be universal for a whole range of neural architectures.

3.3 Model-improving Methods

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

Data augmentation and pruning methods improve the training data to train a model that is robust against a given adversary type. Thus they are inherently linked with the adversarial data generation methods. However, simply training the model on parts of the adversarial evaluation set is not always sufficient, as adversarially robust generalisation increases the sample complexity, and therefore "requires more (training) data" (Schmidt et al., 2018). Wang and Bansal (2018) introduce various improvements to the original ADDSENT algorithm, in order to generate enough training data to obtain robustness for the adversarial evaluation set introduced by Jia and Liang (2017). Geiger et al. (2019) propose a method to estimate the required size of the training set for any given adversarial evaluation set and apply their theory on evaluating the capability of neural networks to learn compositionality.

As an alternative to augmenting training data, Sakaguchi et al. (2019) introduce AFLITE, a theoretically and empirically grounded (Bras et al., 2020) method to automatically detect and remove data points that contribute to arbitrary spurious correlations. Furthermore, we include the application of *adversarial data generation* against a target model when employed during the construction of a new dataset. in crowd-sourcing, where humans act as adversary generators and an entry is only accepted if it triggers a wrong prediction by a trained target model (Nie et al., 2019b; Dua et al., 2019b), or when automatically generating multiple choice alternatives until a target model cannot distinguish between human-written (correct) and automatically generated (wrong) options, called *Adversarial Filtering* (Zellers et al., 2018). Finally, Mishra et al. (2020) combine multiple textual metrics as a measure for data quality and employ adversarial data generation techniques to assist the generation of new dataset resources.

Architecture and Training Procedure Improvements deviate from the idea of data augmentation and seek to train robust models from potentially biased data. These methods include jointly training a model together with one that is designed to exploit correlations and then discarded for inference (Mahabadi et al., 2020; Clark et al., 2019; He et al., 2019), reducing the contribution of "biased" samples to the overall training objective (Schuster et al., 2019; Zhang et al., 2019b), parameter regularisation (Sagawa et al., 2020) and the use of external resources, such as linguistic knowledge (Zhou et al., 2019; Wu et al., 2019) or logic (Minervini and Riedel, 2018).

4 Results

We report the findings of our literature survey in this section. For a full list of investigated datasets and the type(s) of method(s) applied, please refer to Appendix A. Almost half of the surveyed papers (41) are focusing on the RTE task, followed by analysis of the MRC (35) task with 7 and 4 investigating other and

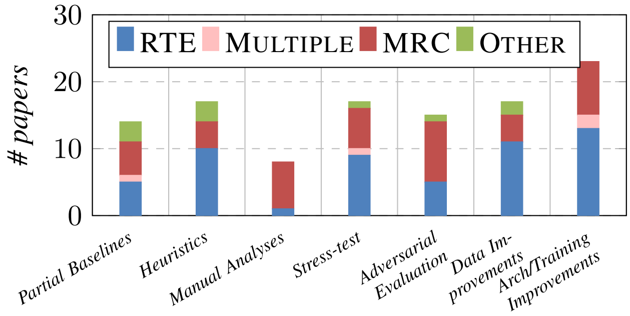

Figure 4: Number of methods per category split by task. As multiple papers report more than one method, the maximum (111) does not add up to the number of surveyed papers (87).

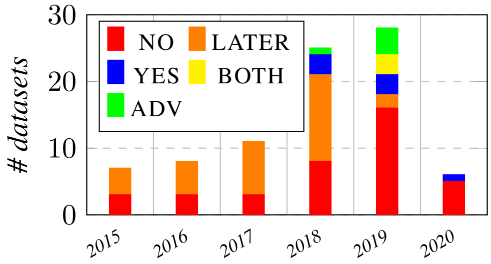

Figure 5: Dataset by publication year with no or any spurious correlations detection methods applied; applied in a later publication; created using adversarial filtering, or both.

multiple tasks, respectively. Looking at the breakdown by type of analysis according to our taxonomy (Figure 4) we see that most approaches concern model-improving methods (c.f. Section 3.2, it is further worth pointing out that every surveyed paper presenting those methods was published in the years 2018-2020. We attribute this popularity to the availability of adversarial evaluation sets such as AddSent (Jia and Liang, 2017) for MRC and HANS (McCoy et al., 2019) for RTE. In fact, the respective source datasets those evaluation sets were derived from, SQuAD and MNLI, are the most utilised resources in the surveyed literature (with 28 and 23 papers investigating or using them). In general, 16 RTE and 33 MRC datasets were analysed or used at least once; we attribute the difference to the existence of various different MRC datasets and the tendency of performing multi-dataset analyses in papers that investigate MRC datasets (Kaushik and Lipton, 2018; Sugawara et al., 2020; Si et al., 2019).

It is worth highlighting that there is little work analysing MRC data 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 or lexical answer type), as the input is a whole paragraph and a question and the expected output is typically a span anywhere in the paragraph ($\mathcal{O}(n)$). For RTE, by contrast, where the input consists of two sentences and the expected output is one of three fixed class labels ($\mathcal{O}(1)$), possible correlations are easier to unveil. Furthermore, there are few (6) MRC "stress-tests" concerning rather broad categories such as prediction consistency (Ribeiro et al., 2019), acquired knowledge (Richardson and Sabharwal, 2019), performance on counterfactual examples (Gardner et al., 2020) or transfer to different datasets (Dua et al., 2019a). Perhaps due to the comparatively higher complexity of the output space, deeper analyses such as the presence of linguistic features and reasoning requirements and the degree to which models acquire them are performed manually (7 papers).

Finally, we report, whether the existence of spurious correlations was investigated in the original or a later publication of a dataset, by applying quantitative methods such as those discussed in Section 3.1: Partial Baselines and Heuristics and Correlations, or whether some sort of adversarial pruning discussed in 3.2: Adversarial Evaluation 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 1.3) 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 advanced correlations in their data (three in 2018 and six in 2019). Furthermore, newly proposed resources are increasingly pruned against state of the art approaches (eight in 2018 and 2019 cumulative). However, for nearly a half (38 out of 85) of the datasets under investigation there is no information about potential spurious correlations yet.

5 Discussion and Conclusion

We present a structured survey of methods that reveal flaws in NLI datasets, methods which show that neural models inherit those correlations or assess their capabilities otherwise, and methods that mitigate this by adversarial training, data augmentation and model architecture or training procedure improvements. Various NLI datasets are reported to contain spurious correlations between input and expected output, might be unsuitable to evaluate some task modality due to dataset design or suffer from quality issues. RTE is a popular target task for these data-centred investigations with more than half of the surveyed papers focusing on it. NLI models, in turn, are shown to exploit those correlations and to rely on superficial lexical cues. Furthermore, they lack generalisation beyond the evaluation set resulting in poor performance on out-of-distribution evaluation sets, generated adversarially or targeted at a specific capability. Efforts to achieve robustness include augmenting the training data with adversarial examples, making use of external resources and modifying the neural network architecture or training objective.

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

- While transformer-based architectures are reported to "improve out-of-distribution robustness" (Hendrycks et al., 2020), partially in line with findings of Kavumba et al. (2019), there is a need for a study that systematically investigates the benefits of type and amount of prior knowledge on neural models' out-of-distribution stress test performance.
- We believe the scientific community will benefit from an application of the quantitative methods that have been presented in this survey to the those NLI datasets that have not been examined for spurious correlations yet (c.f. Appendix B).
- Partial baselines are conceptually simple and cheap to employ for any given task, so we want to incentivise researchers 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 serve as an upper bound for the complexity of the dataset.
- Adapting methods applied to RTE datasets or developing novel methodology to reveal cues and spurious correlations in MRC data is a possible future research direction.
- While RTE is increasingly becoming a proxy 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 shown yet. Additionally, the MRC task requires further capabilities 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 "stress-tests" for MRC models as well, in order to gain a more focussed understanding of their capabilities and limitations.

A noteworthy corollary of the survey is 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 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, is a necessary prerequisite to develop methods as those discussed in this survey, and is therefore a vital step in order to gain empirically grounded understanding of what the current state-of-the-art NLI 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

Asma Ben Abacha, Duy Dinh, and Yassine Mrabet. 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)*, volume 9105, pages 238–242. Springer Verlag.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: BookTest Dataset for Reading Comprehension. *arXiv preprint arXiv 1610.00956*, 10.

Dominika Basaj, Barbara Rychalska, Przemyslaw Biecek, and Anna Wroblewska. 2018. How much should you ask? On the question structure in QA systems. *arXiv preprint arXiv 1809.03734*, 9.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 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*, pages 877–891, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial Filters of Dataset Biases. *arXiv preprint arXiv:2002.04108*, 2.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 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)*, pages 616–622, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jifan Chen and Greg Durrett. 2019a. 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)*, pages 4026–4032, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jifan Chen and Greg Durrett. 2019b. 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)*, pages 4026–4032, Stroudsburg, PA, USA. Association for Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 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)*, volume 4, pages 2358–2367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tiffany Chien and Jugal Kalita. 2020. Adversarial Analysis of Natural Language Inference Systems. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 1–8, 12.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North*, pages 2924–2936, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3944 LNAI, pages 177–190.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–222.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 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)*, pages 1595–1604, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming Question Answering Datasets Into Natural Language Inference Datasets. *arXiv preprint arXiv:1809.02922*, 9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Matt Gardner, and Sameer Singh. 2019a. Comprehensive Multi-Dataset Evaluation of Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 147–153, Stroudsburg, PA, USA, 11. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 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)*, pages 2368–2378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *arXiv preprint arXiv 1704.05179*, 4.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading Failures of Partial-input Baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5533–5538, Stroudsburg, PA, USA, 5. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating NLP Models via Contrast Sets. *arXiv preprint arXiv 2004.02709*, 4.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 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)*, pages 4484–4494, Stroudsburg, PA, USA, 11. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 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)*, pages 650–655, Stroudsburg, PA, USA, 5. Association for Computational Linguistics.

Quentin Grail, Julien Perez, and Tomi Silander. 2018. Adversarial Networks for Machine Reading. *TAL Traitement Automatique des Langues*, 59(2):77–100.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 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)*, pages 1930–1940, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. *arXiv preprint arXiv 2005.06676*, 5.

He He, Sheng Zha, and Haohan Wang. 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)*, pages 132–142, Stroudsburg, PA, USA, 8. Association for Computational Linguistics.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained Transformers Improve Out-of-Distribution Robustness. *arXiv preprint arXiv 2004.06100*, 4.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. *Advances in Neural Information Processing Systems*, 2015-Janua:1693–1701, 6.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering. *arXiv preprint arXiv 2005.05257*, 5.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 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)*, pages 2391–2401, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 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)*, pages 4127–4140, Stroudsburg, PA, USA, 9. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 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*, pages 2726–2736, Stroudsburg, PA, USA, 6. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 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)*, pages 2567–2577, Stroudsburg, PA, USA, 9. Association for Computational Linguistics.

Yimin Jing, Deyi Xiong, and Zhen Yan. 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)*, pages 2452–2462, Stroudsburg, PA, USA, 10. Association for Computational Linguistics.

Tomasz Jurczyk, Michael Zhai, and Jinho D. Choi. 2016. SelQA: A New Benchmark for Selection-based Question Answering. *Proceedings - 2016 IEEE 28th International Conference on Tools with Artificial Intelligence, ICTAI 2016*, pages 820–827, 6.

Sanjay Kamath, Brigitte Grau, and Yue Ma. 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*, pages 72–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. 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)*, volume 1, pages 2418–2428, Stroudsburg, PA, USA, 5. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 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*, pages 5010–5015, Stroudsburg, PA, USA. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. In *International Conference on Learning Representations*, 9.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42. Association for Computational Linguistics (ACL), 11.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the First Sentence: Position Bias in Question Answering. *arXiv preprint arXiv 2004.14602*, 4.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 3.

Alice Lai and Julia Hockenmaier. 2014. Illinois-LH: A Denotational and Distributional Approach to Semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334, Stroudsburg, PA, USA, 6. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering. *arXiv preprint arXiv 1607.06275*, 7.

Yichan Liang, Jianheng Li, and Jian Yin. 2019. A New Multi-choice Reading Comprehension Dataset for Curriculum Learning. In *Proceedings of Machine Learning Research*, volume 101, pages 742–757. International Machine Learning Society (IMLS), 3.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning Over Paragraph Effects in Situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pengyuan Liu, Chengyu Du, Shuofeng Zhao, and Chenghao Zhu. 2019a. Emotion Action Detection and Emotion Inference: the Task and Dataset. *arXiv preprint arXiv 1903.06901*, 3.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019b. Neural Machine Reading Comprehension: Methods and Trends. *Applied Sciences*, 9(18):3698, 9.

Kai Liu, Xin Liu, An Yang, Jing Liu, Jinsong Su, Sujian Li, and Qiaoqiao She. 2020. A Robust Adversarial Training Approach to Machine Reading Comprehension. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8392–8400.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-End Bias Mitigation by Modelling Biases in Corpora. *arXiv preprint arXiv 1909.06321*, 9.

Gengchen Mai, Krzysztof Janowicz, Cheng He, Sumang Liu, and Ni Lao. 2018. POIReviewQA: A semantically enriched POI retrieval and question answering dataset. In *Proceedings of the 12th Workshop on Geographic Information Retrieval, GIR 2018*, pages 1–2, New York, New York, USA, 11. Association for Computing Machinery, Inc.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 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*, pages 3428–3448, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv preprint arXiv:1908.09635*, 8.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 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*, pages 2381–2391, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 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)*, volume 1, pages 1725–1735, Stroudsburg, PA, USA, 5. Association for Computational Linguistics.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional Questions Do Not Necessitate Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Stroudsburg, PA, USA, 6. Association for Computational Linguistics.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. *arXiv preprint arXiv 2004.11999*, 4.

Pasquale Minervini and Sebastian Riedel. 2018. Adversarially Regularising Neural NLI Models to Integrate Logical Background Knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 65–74, Stroudsburg, PA, USA, 8. Association for Computational Linguistics.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020. DQI: Measuring Data Quality in NLP. *arXiv preprint arXiv 2005.00816*, 5.

Arindam Mitra, Ishan Shrivastava, and Chitta Baral. 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*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 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*, pages 839–849, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the Model Understand the Question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1896–1906, Stroudsburg, PA, USA, 5. Association for Computational Linguistics.

James Mullenbach, Jonathan Gordon, Nanyun Peng, and Jonathan May. 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)*, pages 6051–6057, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, page 2340–2353, Santa Fe, New Mexico, USA, 8. Association for Computational Linguistics.

Mao Nakanishi, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2018. Answerable or Not: Devising a Dataset for Extending Machine Reading Comprehension. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 973–983.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019a. Analyzing Compositionality-Sensitivity of NLI Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874, 7.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019b. Adversarial NLI: A New Benchmark for Natural Language Understanding. *arXiv preprint arXiv:1910.14599*, 10.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. *arXiv preprint arXiv 2005.00242*, 5.

Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Stroudsburg, PA, USA, 7. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 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*, pages 2357–2368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mithun Paul Panenghat, Sandeep Suntwal, Faiz Rafique, Rebecca Sharp, and Mihai Surdeanu. 2020. Towards the Necessity for Debiasing Natural Language Inference Datasets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, page 6883–6888. European Language Resources Association.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 11.

Oskar Pfungst and Carl Leo Rahn. 1911. *Clever Hans (the horse of Mr. Von Osten) a contribution to experimental animal and human psychology,*, volume 8. H. Holt and company, New York,.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hemant Pugaliya, James Route, Kaixin Ma, Yixuan Geng, and Eric Nyberg. 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*, pages 125–136, Stroudsburg, PA, USA, 11. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683*, 10.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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)*, volume 1, pages 856–865, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are Red Roses Red? Evaluating Consistency of Question-Answering Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kyle Richardson and Ashish Sabharwal. 2019. What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge. *arXiv preprint arXiv:1912.13337*, 12.

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2019. Probing Natural Language Inference Models through Semantic Fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8722–8731. AAAI Press.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Stroudsburg, PA, USA, 7. Association for Computational Linguistics.

Barbara Rychalska, Dominika Basaj, Anna Wróblewska, and Przemyslaw Biecek. 2018. 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*, pages 322–324, Stroudsburg, PA, USA, 9. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *International Conference on Learning Representations*, 11.

Tanik Saikh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. ScholarlyRead: A New Dataset for Scientific Article Reading Comprehension. In *Proceedings of The 12th Language Resources and Evaluation Conference*, page 5498–5504. European Language Resources Association.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *arXiv preprint arXiv:1907.10641*, 7.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 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)*, pages 1975–1985, Stroudsburg, PA, USA, 5. Association for Computational Linguistics.

Viktor Schlegel, Marco Valentino, André Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A Framework for Evaluation of Machine Reading Comprehension Gold Standards. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 3.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Google Brain. 2018. Adversarially Robust Generalization Requires More Data. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, pages 5014–5026.

Martin Schmitt and Hinrich Schütze. 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*, pages 902–914, Stroudsburg, PA, USA, 6. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 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)*, pages 3417–3423, Stroudsburg, PA, USA, 8. Association for Computational Linguistics.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 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)*, pages 15–25, Stroudsburg, PA, USA. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *International Conference on Learning Representations*.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does BERT Learn from Multiple-Choice Reading Comprehension Datasets? *arXiv preprint arXiv:1910.12391*, 10.

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2020. Benchmarking Robustness of Machine Reading Comprehension Models. *arXiv preprint arXiv 2004.14004*, 4.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 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)*, pages 4505–4514, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. There is Strength in Numbers: Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. *arXiv preprint arXiv 2004.07790*, 4.

Janez Starc and Dunja Mladenić. 2017. Constructing a Natural Language Inference dataset using generative neural networks. *Computer Speech & Language*, 46:94–112, 11.

Petros Stavropoulos, Dimitris Pappas, Ion Androutsopoulos, and Ryan McDonald. 2020. BIOMRC: A Dataset for Biomedical Machine Reading Comprehension. *arXiv preprint arXiv 2005.06376*, 5.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 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)*, pages 806–817, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017b. Prerequisite Skills for Reading Comprehension: Multi-Perspective Analysis of MCTest Datasets and Systems. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 3089–3096, 2.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What Makes Reading Comprehension Questions Easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Shawn Tan, Yikang Shen, Chin-wei Huang, and Aaron Courville. 2019. Investigating Biases in Textual Entailment Datasets. *arXiv preprint arXiv 1906.09635*, 6.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 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)*, pages 6075–6084, Stroudsburg, PA, USA, 9. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 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)*, pages 2944–2953, Stroudsburg, PA, USA, 11. Association for Computational Linguistics.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 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)*, pages 3380–3385, Stroudsburg, PA, USA, 11. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Measuring and Reducing Non-Multifact Reasoning in Multi-hop Question Answering. *arXiv preprint arXiv 2005.00789*, 5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 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 Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Stroudsburg, PA, USA, 11. Association for Computational Linguistics.

Yicheng Wang and Mohit Bansal. 2018. Robust Machine Comprehension Models via Adversarial Training. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2 (Short P:575–581.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkši´cmrkši´c, Milica Gaši´cgaši´c, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 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*, volume 1, pages 438–449. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pages 1112–1122, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhijing Wu and Hua Xu. 2020. Improving the robustness of machine reading comprehension model with hierarchical knowledge and auxiliary unanswerability prediction. *Knowledge-Based Systems*, 203:106075, 9.

Bowen Wu, Haoyang Huang, Zongsheng Wang, Qihang Feng, Jingsong Yu, and Baoxun Wang. 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*, pages 53–57, Stroudsburg, PA, USA, 11. Association for Computational Linguistics.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A Social Media Focused Question Answering Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Stroudsburg, PA, USA. Association for Computational Linguistics.

Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. 2019. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *arXiv preprint arXiv:1909.08072*, 9.

Yadollah Yaghoobzadeh, Remi Tachet, T. J. Hazen, and Alessandro Sordoni. 2019. Robust Natural Language Inference Models with Example Forgetting. *arXiv preprint arXiv 1911.03861*, 11.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can Neural Networks Understand Monotonicity Reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Stroudsburg, PA, USA, 6. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 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)*, pages 250–255, Stroudsburg, PA, USA, 4. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Yatskar. 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)*, pages 2318–2323, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson D'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and Evaluating General Linguistic Intelligence. *arXiv preprint arXiv:1901.11373*, 1.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations*, 2.

Fengcheng Yuan, Zheng Lin, Yue Geng, Weiping Wang, and Gang Shi. 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)*, pages 752–759. IEEE, 12.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Stroudsburg, PA, USA, 6. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Stroudsburg, PA, USA, 5. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv preprint arXiv:1810.12885*.

Guanhua Zhang, Bing Bai, Jian Liang, Kun Bai, Shiyu Chang, Mo Yu, Conghui Zhu, and Tiejun Zhao. 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*, pages 4418–4429, Stroudsburg, PA, USA, 5. Association for Computational Linguistics.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2019b. Mitigating Annotation Artifacts in Natural Language Inference Datasets to Improve Cross-dataset Generalization Ability. *arXiv preprint arXiv 1909.04242*, 9.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2019c. Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey. *arXiv preprint arXiv:1901.06796*, 1.

Xiang Zhou and Mohit Bansal. 2020. Towards Robustifying NLI Models Against Lexical Dataset Biases. *arXiv preprint arXiv 2005.04732*, 5.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2019. Robust Reading Comprehension with Linguistic Constraints via Posterior Regularization. *arXiv preprint arXiv:1911.06948*, 11.

A Detailed Survey Results

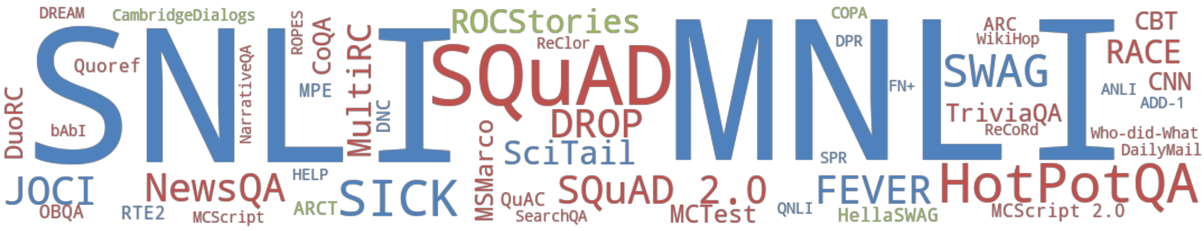

Figure 6: Word cloud with investigated RTE, MRC and other datasets. Size proportional to the number of surveyed papers investigating the dataset.

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
DROP	Stress-test	(Gardner et al., 2020; Dua et al., 2019a)
	Manual Analyses	(Schlegel et al., 2020)
	Adversarial Evaluation	(Dua et al., 2019b)
Quoref	Stress-test	(Gardner et al., 2020; Dua et al., 2019a)
ROPES	Stress-test	(Gardner et al., 2020; Dua et al., 2019a)
BoolQ	Stress-test	(Gardner et al., 2020)
MCTACO	Stress-test	(Gardner et al., 2020)
ADD-1	Arch/Training Improvements	(Belinkov et al., 2019; Stacey et al., 2020)
	Heuristics	(Poliak et al., 2018)
	Partial Baselines	(Poliak et al., 2018)
DPR	Arch/Training Improvements	(Belinkov et al., 2019; Stacey et al., 2020)
	Heuristics	(Poliak et al., 2018)
	Partial Baselines	(Poliak et al., 2018)
FN+	Arch/Training Improvements	(Belinkov et al., 2019; Stacey et al., 2020)
	Heuristics	(Poliak et al., 2018)
	Partial Baselines	(Poliak et al., 2018)
JOCI	Arch/Training Improvements	(Belinkov et al., 2019; Stacey et al., 2020; Zhang et al., 2019b)
	Manual Analyses	(Pavlick and Kwiatkowski, 2019)
	Heuristics	(Poliak et al., 2018)
	Partial Baselines	(Poliak et al., 2018)
MNLI	Arch/Training Improvements	(Belinkov et al., 2019; Mahabadi et al., 2020; Stacey et al., 2020; Zhou and Bansal, 2020; Yaghoobzadeh et al., 2019; Mitra et al., 2020; Minervini and Riedel, 2018; He et al., 2019; Wang et al., 2019; Sagawa et al., 2020; Zhang et al., 2019b; Clark et al., 2019)
	Adversarial Evaluation	(Han et al., 2020; Chien and Kalita, 2020; Nie et al., 2019a)
	Data Improvements	(Panenghat et al., 2020; Min et al., 2020; Zhou and Bansal, 2020; Mitra et al., 2020)
	Heuristics	(Tan et al., 2019; Nie et al., 2019a; Zhang et al., 2019a; Bras et al., 2020; Poliak et al., 2018; McCoy et al., 2019; Gururangan et al., 2018)
	Manual Analyses	(Pavlick and Kwiatkowski, 2019)

	Stress-test	(Naik et al., 2018; Liu et al., 2019a; Nie et al., 2019a; Richardson et al., 2019; Glockner et al., 2018; McCoy et al., 2019)
	Partial Baselines	(Nie et al., 2019a; Poliak et al., 2018; Gururangan et al., 2018)
MPE	Arch/Training Improvements	(Belinkov et al., 2019; Stacey et al., 2020)
	Heuristics	(Poliak et al., 2018)
	Partial Baselines	(Poliak et al., 2018)
SICK	Arch/Training Improvements	(Belinkov et al., 2019; Stacey et al., 2020; Wang et al., 2019; Zhang et al., 2019b)
	Heuristics	(Zhang et al., 2019a; Poliak et al., 2018)
	Partial Baselines	(Lai and Hockenmaier, 2014; Poliak et al., 2018)
SNLI	Arch/Training Improvements	(Belinkov et al., 2019; Mahabadi et al., 2020; Stacey et al., 2020; Jia et al., 2019; Mitra et al., 2020; Minervini and Riedel, 2018; He et al., 2019; Zhang et al., 2019b)
	Adversarial Evaluation	(Mishra et al., 2020; Nie et al., 2019a; Sanchez et al., 2018)
	Heuristics	(Mishra et al., 2020; Tan et al., 2019; Nie et al., 2019a; Zhang et al., 2019a; Bras et al., 2020; Rudinger et al., 2017; Poliak et al., 2018; Gururangan et al., 2018)
	Manual Analyses	(Pavlick and Kwiatkowski, 2019)
	Data Improvements	(Kaushik et al., 2020; Mitra et al., 2020; Kang et al., 2018)
	Stress-test	(Kaushik et al., 2020; Nie et al., 2019a; Richardson et al., 2019; Glockner et al., 2018)
	Partial Baselines	(Nie et al., 2019a; Feng et al., 2019; Poliak et al., 2018; Gururangan et al., 2018)
SPR	Arch/Training Improvements	(Belinkov et al., 2019; Stacey et al., 2020)
	Heuristics	(Poliak et al., 2018)
	Partial Baselines	(Poliak et al., 2018)
SciTail	Arch/Training Improvements	(Belinkov et al., 2019; Stacey et al., 2020)
	Heuristics	(Poliak et al., 2018)
	Partial Baselines	(Poliak et al., 2018)
	Stress-test	(Glockner et al., 2018)
SQuAD	Arch/Training Improvements	(Yuan et al., 2019; Liu et al., 2020; Ko et al., 2020; Wu and Xu, 2020; Min et al., 2018; Zhou et al., 2019; Wu et al., 2019; Clark et al., 2019)
	Manual Analyses	(Sugawara et al., 2017a; Sugawara et al., 2018; Pugaliya et al., 2019)
	Heuristics	(Sugawara et al., 2018; Ko et al., 2020)
	Adversarial Evaluation	(Basaj et al., 2018; Mudrakarta et al., 2018; Rychalska et al., 2018; Wallace et al., 2019; Jia and Liang, 2017)
	Stress-test	(Nakanishi et al., 2018; Dua et al., 2019a; Liu et al., 2019a; Ribeiro et al., 2019)
	Data Improvements	(Nakanishi et al., 2018; Wang and Bansal, 2018)
	Partial Baselines	(Sugawara et al., 2020; Kaushik and Lipton, 2018)
MCTest	Manual Analyses	(Sugawara et al., 2017b; Sugawara et al., 2017a; Sugawara et al., 2018)
	Heuristics	(Sugawara et al., 2018)
	Partial Baselines	(Si et al., 2019; Sugawara et al., 2020)
	Adversarial Evaluation	(Si et al., 2019)

MSMarco	Manual Analyses Heuristics	(Sugawara et al., 2017a; Sugawara et al., 2018; Schlegel et al., 2020; Pugaliya et al., 2019) (Sugawara et al., 2018)
NewsQA	Manual Analyses Heuristics Arch/Training Improvements Stress-test	(Sugawara et al., 2017a; Sugawara et al., 2018; Schlegel et al., 2020) (Sugawara et al., 2018) (Min et al., 2018) (Dua et al., 2019a)
QA4MRE	Manual Analyses	(Sugawara et al., 2017a)
Who-did-What	Manual Analyses Partial Baselines	(Sugawara et al., 2017a) (Kaushik and Lipton, 2018)
ARC	Heuristics Manual Analyses Stress-test	(Sugawara et al., 2018) (Sugawara et al., 2018) (Richardson and Sabharwal, 2019)
TriviaQA	Heuristics Manual Analyses Arch/Training Improvements	(Sugawara et al., 2018) (Sugawara et al., 2018) (Min et al., 2018; Clark et al., 2019)
WikiHop	Heuristics Manual Analyses Partial Baselines	(Sugawara et al., 2018) (Sugawara et al., 2018) (Chen and Durrett, 2019b)
Narrative-QA	Heuristics Manual Analyses Stress-test	(Sugawara et al., 2018) (Sugawara et al., 2018) (Dua et al., 2019a)
RACE	Heuristics Manual Analyses Adversarial Evaluation Partial Baselines	(Sugawara et al., 2018) (Sugawara et al., 2018) (Si et al., 2020; Si et al., 2019) (Si et al., 2019; Sugawara et al., 2020)
MCScript	Heuristics Manual Analyses Partial Baselines Adversarial Evaluation	(Sugawara et al., 2018) (Sugawara et al., 2018) (Si et al., 2019) (Si et al., 2019)
FEVER	Data Improvements Adversarial Evaluation Heuristics Arch/Training Improvements	(Panenghat et al., 2020; Schuster et al., 2019) (Thorne et al., 2019) (Schuster et al., 2019) (Schuster et al., 2019)
HotPotQA	Partial Baselines Stress-test Heuristics Manual Analyses Adversarial Evaluation Data Improvements Arch/Training Improvements	(Trivedi et al., 2020; Sugawara et al., 2020; Min et al., 2019; Chen and Durrett, 2019b) (Trivedi et al., 2020) (Trivedi et al., 2020) (Schlegel et al., 2020; Pugaliya et al., 2019) (Jiang and Bansal, 2019) (Jiang and Bansal, 2019) (Jiang and Bansal, 2019)
MultiRC	Manual Analyses	(Schlegel et al., 2020)

	Partial Baselines	(Sugawara et al., 2020)
ReCoRd	Manual Analyses	(Schlegel et al., 2020)
DNC	Manual Analyses	(Pavlick and Kwiatkowski, 2019)
RTE2	Manual Analyses	(Pavlick and Kwiatkowski, 2019)
CBT	Arch/Training Improvements	(Grail et al., 2018)
	Partial Baselines	(Kaushik and Lipton, 2018)
Cambridge-Dialogs	Arch/Training Improvements	(Grail et al., 2018)
DuoRC	Stress-test	(Dua et al., 2019a)
	Partial Baselines	(Sugawara et al., 2020)
SQuAD 2.0	Stress-test	(Dua et al., 2019a)
	Partial Baselines	(Sugawara et al., 2020)
	Manual Analyses	(Yatskar, 2019)
HELP	Data Improvements	(Yanaka et al., 2019b)
SWAG	Partial Baselines	(Trichelair et al., 2019; Sugawara et al., 2020)
	Adversarial Evaluation	(Zellers et al., 2018; Zellers et al., 2019)
DREAM	Partial Baselines	(Si et al., 2019)
	Adversarial Evaluation	(Si et al., 2019)
MCScript 2.0	Partial Baselines	(Si et al., 2019)
	Adversarial Evaluation	(Si et al., 2019)
QNLI	Heuristics	(Bras et al., 2020)
OBQA	Stress-test	(Richardson and Sabharwal, 2019)
ReClor	Heuristics	(Yu et al., 2020)
CoQA	Partial Baselines	(Sugawara et al., 2020)
	Manual Analyses	(Yatskar, 2019)
CNN	Manual Analyses	(Chen et al., 2016)
	Partial Baselines	(Kaushik and Lipton, 2018)
DailyMail	Manual Analyses	(Chen et al., 2016)
ROCStories	Heuristics	(Cai et al., 2017)
	Partial Baselines	(Cai et al., 2017; Schwartz et al., 2017)
HellaSWAG	Adversarial Evaluation	(Zellers et al., 2019)
SearchQA	Manual Analyses	(Pugaliya et al., 2019)
COPA	Heuristics	(Kavumba et al., 2019)
	Stress-test	(Kavumba et al., 2019)
ARCT	Heuristics	(Niven and Kao, 2019)
	Adversarial Evaluation	(Niven and Kao, 2019)
ANLI	Adversarial Evaluation	(Nie et al., 2019b)
QuAC	Manual Analyses	(Yatskar, 2019)
bAbI	Partial Baselines	(Kaushik and Lipton, 2018)

The following table shows those 36 datasets from Figure 5 broken down by year, where no quantitative methods to describe possible spurious correlations have been applied yet:

Year	**Dataset**
2015	DailyMail (Hermann et al., 2015), MedlineRTE (Abacha et al., 2015), WikiQA (Yang et al., 2015)
2016	SelQA (Jurczyk et al., 2016), WebQA (Li et al., 2016), BookTest (Bajgar et al., 2016)

2017	CambridgeDialogs (Wen et al., 2017), SearchQA (Dunn et al., 2017), GANNLI (Starc and Mladenić, 2017)
2018	OBQA (Mihaylov et al., 2018), QuAC (Choi et al., 2018), MedHop (Welbl et al., 2018), BioASQ (Kamath et al., 2018), PoiReviewQA (Mai et al., 2018), emrQA (Pampari et al., 2018), ProPara (Dalvi et al., 2018), ReCoRd (Zhang et al., 2018)
2019	BoolQ (Clark et al., 2019), MCTACO (Zhou et al., 2019), ROPES (Lin et al., 2019), SherLIiC (Schmitt and Schütze, 2019), CLUTRR (Sinha et al., 2019), BiPaR (Jing et al., 2019), NaturalQ (Kwiatkowski et al., 2019), CosmosQA (Huang et al., 2019), VGNLI (Mullenbach et al., 2019), PubMedQA (Jin et al., 2019), WIQA (Tandon et al., 2019), TWEET-QA (Xiong et al., 2019), HEAD-QA (Vilares and Gómez-Rodríguez, 2019), RACE-C (Liang et al., 2019), CEAC (Liu et al., 2019a), HELP (Yanaka et al., 2019b)
2020	QuAIL (Rogers et al., 2020), ScholarlyRead (Saikh et al., 2020), BioMRC (Stavropoulos et al., 2020), TORQUE (Ning et al., 2020), SARA (Holzenberger et al., 2020)

B 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 6 by a Google Scholar search using the queries shown in Table 3. 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.

```
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"
```

Table 3: Google Scholar Queries for the extended dataset corpus