Teach Me to Explain: A Review of Datasets for Explainable NLP

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Ana Marasović*
Allen Institute for AI
University of Washington
anam@allenai.org

Abstract

Explainable NLP (ExNLP) has increasingly focused on collecting human-annotated explanations. These explanations are used downstream in three ways: as data augmentation to improve performance on a predictive task, as a loss signal to train models to produce explanations for their predictions, and as a means to evaluate the quality of model-generated explanations. In this review, we identify three predominant classes of explanations (highlights, free-text, and structured), organize the literature on annotating each type, point to what has been learned to date, and give recommendations for collecting ExNLP datasets in the future.

1 Introduction

Interpreting supervised machine learning models by analyzing *local explanations*—justifications of their individual predictions[1]—is important for debugging, protecting sensitive information, understanding model robustness, and increasing warranted trust in a stated contract (Doshi-Velez and Kim, 2017; Molnar, 2019; Jacovi et al., 2021). Explainable NLP (ExNLP) researchers rely on datasets that contain human justifications for the true label (overviewed in Tables 3–5). Human justifications are used to aid models with additional training supervision (*learning-from-explanations*; Zaidan et al., 2007), to train interpretable models that explain their own predictions (Camburu et al., 2018), and to evaluate plausibility of explanations by measuring their agreement with human-annotated explanations (DeYoung et al., 2020a).

Dataset collection is the most under-scrutinized component of the machine learning pipeline (Paritosh, 2020)—it is estimated that 92% of machine learning practitioners encounter data cascades, or downstream problems resulting from poor data quality (Sambasivan et al., 2021). It is important to constantly evaluate data collection practices critically and establish common and standardized methodologies (Bender and Friedman, 2018; Gebru et al., 2018; Jo and Gebru, 2020; Ning et al., 2020; Paullada et al., 2020). Even highly influential datasets that are widely adopted in research undergo such inspections. For instance, Recht et al. (2019) tried to reproduce the ImageNet test set by carefully following the instructions in the paper (Deng et al., 2009), but multiple models do not perform well on the replicated test set. We expect that such examinations are particularly valuable when many related datasets are released contemporaneously and independently in a short period of time, as is the case with ExNLP datasets.

This survey aims to review and summarize the literature on collecting ExNLP datasets, highlight what has been learned to date, and give recommendations for future ExNLP dataset construction. It complements other explainable AI (XAI) surveys and critical retrospectives that focus on definitions (Gilpin et al., 2018; Miller, 2019; Clinciu and Hastie, 2019; Barredo Arrieta et al., 2020; Jacovi and Goldberg, 2020), methods (Biran and Cotton, 2017; Ras et al., 2018; Guidotti et al., 2019), evaluation (Hoffman et al., 2018; Yang et al., 2019), or a combination of these topics (Lipton, 2018; Doshi-Velez and Kim, 2017; Adadi and Berrada, 2018; Murdoch et al., 2019; Verma et al., 2020; Burkart and Huber, 2021), but not on datasets. Kotonya and Toni (2020a) review datasets and methods for explaining automated fact-checking; we are the first to review datasets across all NLP tasks.

To organize the literature, we first define our scope (§2), sort out ExNLP terminology (§3), and present an overview of existing ExNLP datasets

*Equal contributions.

[1]As opposed to global explanation, where one explanation is given for an entire system (Molnar, 2019).

(§4).[2] Then, we present what can be learned from existing ExNLP datasets. Specifically, §5 discusses the typical process of collecting highlighted elements in the inputs as explanations, and discrepancies between this process and that of evaluating automatic methods. In the same section, we also draw attention to how assumptions made in free-text explanation collection influence their modeling, and call for better documenting of ExNLP data collection. In §6, we illustrate that not all template-like free-text explanations are unwanted, and call for embracing the structure of an explanation when appropriate. In §7.1, we present a proposal for controling quality in free-text explanation collection. Finally, §8 gathers recommendations from related subfields to further reduce data artifacts by increasing diversity of collected explanations.

2 Survey Scope

An explanation can be described as a "three-place predicate: *someone* explains *something* to *someone*" (Hilton, 1990). But what is the *something* being explained in ExNLP? We identify two types of ExNLP dataset papers, based on their answer to this question: those which explain *human decisions* (task labels), and those which explain *observed events or phenomena* in the world. We illustrate these classes in Figure 1.

The first class, explaining *human decisions*, has a long history and strong association with the term "explanation" in NLP. In this class, explanations are not the task labels themselves, but rather explain some task label for a given input (see examples in Figure 1 and Table 1). One distinguishing factor in their collection is that they are often collected by asking "why" questions, such as "why is [input] assigned [label]?". The underlying assumption is that the explainer believes the assigned label to be correct or at least likely, either because they selected it, or another annotator did (discussed further in §7.3). Resulting datasets contain instances that are tuples of the form (input, label, explanation), where each explanation is tied to an input-label pair. Of course, collecting explanations of human decisions can help train systems to model them. This class is thus directly suited to the three ExNLP goals: they serve as (i) additional training supervision to predict labels from inputs, (ii) training supervision for models to explain their label predictions, and (iii) an evaluation set for the quality of model-produced explanations of label predictions. We define our survey scope as focusing on this class of explanations.

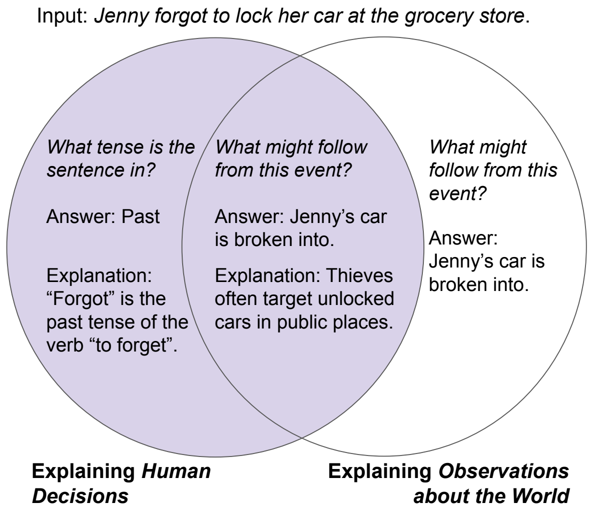


Figure 1: A description of the two classes of ExNLP datasets (§2). For some input, an explanation can explain a human-provided answer, the answer (left and middle) or something about the world (middle and right). The shaded area is our survey scope. These classes can be used to teach and evaluate models on different skill sets: their ability to justify their predictions (left), reason about the world (right), or do both (middle).

The second class falls under commonsense reasoning—this type of explanation is most commonly referred to as *commonsense inference* about the world. While most datasets that can be framed as explaining something about the world do not use the term "explanation" (Zellers et al., 2018; Sap et al., 2019b; Zellers et al., 2019b; Fan et al., 2019; Bisk et al., 2020; Forbes et al., 2020), a few recent ones do, e.g., ART (Bhagavatula et al., 2020) and GLUCOSE (Mostafazadeh et al., 2020). These datasets contain explanations as instance-labels themselves, rather than as justifications for existing labels. For example, ART collects plausible hypotheses and GLUCOSE collects certain or likely statements to explain how sentences in a story can follow other sentences. They thus produce tuples of the form (input, explanation), where the input is an event or observation.

The two classes are not mutually exclusive. For example, SBIC (Sap et al., 2019a) contains

[2] We created a living version of the tables as a website, where pull requests and issues can be opened to add and update datasets: `https://exnlpdatasets.github.io`

Instance	**Explanation**
Premise: A shirtless man wearing white shorts. *Hypothesis:* A man in white shorts is running on the sidewalk. *Label:* neutral	`(highlight)` *Premise:* A shirtless man wearing white shorts. *Hypothesis:* A man in white shorts is running on the sidewalk. *Label*: neutral
	`(free-text)` Just because a man is wearing shorts does not mean he is running on the sidewalk.
Question: Who sang the theme song from Russia With Love? *Paragraph:* John Barry, arranger of Monty Norman's "James Bond Theme"...would be the dominant Bond series composer for most of its history and the inspiration for fellow series composer, David Arnold (who uses cues from this soundtrack in his own...). The theme song was composed by Lionel Bart of Oliver! fame and sung by Matt Monro. *Answer:* Matt Monro	`(structured)` *Sentence selection:* The theme song was composed by Lionel Bart of Oliver! fame and sung by Matt Monro. *Referential equality:* "the theme song from russia with love" (extracted from question) = "The theme song" (extracted from paragraph) *Entailment:* X was composed by Lionel Bart of Oliver! fame and sung by ANSWER. $\vdash$ ANSWER sung X

Table 1: Examples of explanation types discussed in §3. The first two rows show a highlight and free-text explanation for an E-SNLI instance (Camburu et al., 2018). The last row shows a structured explanation from QED for a NATURALQUESTIONS instance (Lamm et al., 2020).

human-annotated labels and justifications indicating whether or not social media posts contain social bias. While this clearly follows the form of the first class because it provides explanations for human-assigned bias labels, it also explains observed phenomena in the world: the justification for the phrase "We shouldn't lower our standards just to hire more women" receiving a bias label is "[it] implies women are less qualified." Other examples of datasets in both classes include predicting future events in videos (VLEP; Lei et al., 2020) and answering commonsense questions about images (VCR; Zellers et al., 2019a). Both collect observations about a real-world setting as task labels with rationales explaining why the observations are correct.

This second class is potentially unbounded because many machine reasoning datasets could be re-framed as explanation under this definition. For example, one could consider labels in the PIQA dataset (Bisk et al., 2020) to be explanations, even though they do not explain human decisions. For example, for the question "how do I separate an egg yolk?," the PIQA gold label is "Squeeze the water bottle and press it against the yolk. Release, which creates suction and lifts the yolk." While this instruction is an explanation of how to separate an egg yolk, an explanation for the *human decision* would require a further justification of the given response, e.g., "Separating the egg requires pulling the yolk out. Suction provides the force to do this." We do not survey datasets that only fall into this broader, unbounded class such as PIQA, ART, and GLUCOSE, leaving discussion of how best to define this class and its utility for the goals of EXNLP to future work.

3 Explainability Lexicon

The first step in collecting human explanations is to decide in what format annotators will give their justifications. We identify three types: *highlights*, *free-text*, and *structured* explanations. An example of each type is given in Table 1. Since a consensus on terminology has not yet been reached, we describe each type below.

To define highlights, we turn to Lei et al. (2016) who popularized *extractive rationales*, i.e., subsets of the input tokens that satisfy two properties: (i) *compactness*, the selected input tokens are short and coherent, and (ii) *sufficiency*, the selected tokens suffice for prediction as a substitute of the original text. Yu et al. (2019) introduce a third property, (iii) *comprehensiveness*, that all the evidence that supports the prediction is selected, not just a sufficient set—i.e., the tokens that are *not* selected do *not* predict the label.[3] Lei et al. (2016) formalize the task of first selecting the parts of the input instance to be the extractive explanation, and then making a prediction based only on the extracted tokens. Since then, this task and the term "extractive rationale" have been associated. Jacovi and Goldberg (2021) argue against the term, because "rationalization" implies human intent, while the task Lei et al. (2016) propose does

[3] Henceforth, "sufficient" and "comprehensive" always refer to these definitions.

Instance with Highlight	Highlight Type Clarification
Review: this film is extraordinarily horrendous and I'm not going to waste any more words on it. *Label*: negative	(¬comprehensive) *Review:* this film is [redacted] and I'm not going to waste any more words on it.
Review: claire danes , giovanni ribisi , and omar epps make a likable trio of protagonists , but they 're just about the only palatable element of the mod squad , a lame - brained big - screen version of the 70s tv show .\nthe story has all the originality of a block of wood \n(well , it would if you could decipher it) , the characters are all blank slates , and scott silver 's perfunctory action sequences are as cliched as they come .\nby sheer force of talent , the three actors wring marginal enjoyment from the proceedings whenever they 're on screen , but the mod squad is just a second - rate action picture with a first - rate cast. *Label:* negative	(comprehensive) *Review:* claire danes , giovanni ribisi , and omar epps make a likable trio of protagonists , but they 're just about the only palatable element of the mod squad , [redacted] of the 70s tv show .\nthe story [redacted] \n(well , it would if you could decipher it) , the characters are all blank slates , and scott silver 's perfunctory action sequences are [redacted] \nby sheer force of talent , the three actors wring marginal enjoyment from the proceedings whenever they 're on screen , but the mod squad is just a [redacted] with a first - rate cast.
Premise: A shirtless man wearing white shorts. *Hypothesis:* A man in white shorts is running on the sidewalk. *Label*: neutral	(¬sufficient) *Premise:* [redacted] *Hypothesis:* [redacted] man [redacted] running on the sidewalk.

Table 2: Examples of highlight properties discussed in §3 and §5. The highlight for the first example, a negative movie review (Zaidan et al., 2007), is non-comprehensive because its complement in col. 2 is also predictive of the label, unlike the highlight for the second movie review (DeYoung et al., 2020a). These two highlights are sufficient (not illustrated in col. 2) because they are predictive of the negative label, in contrast to the highlight for an E-SNLI premise-hypothesis pair (Camburu et al., 2018) in the last row, which is not sufficient because it is not predictive of the neutral label on its own.

not rely on, or aim to model, human justifications. As an alternative, they propose the term *highlights* and the verbs "highlighting" or "extracting highlights" to avoid inaccurately attributing human-like social behavior to AI systems. Fact-checking, medical claim verification, and multi-document question answering (QA) research often refers to highlights as *evidence*—a part of the source that refutes or supports the claim.

To reiterate, **highlights** are subsets of the input elements (either words, subwords, or full sentences) that are compact and sufficient to explain a prediction. If they are also comprehensive, we call them *comprehensive highlights*. Although the community has settled on criteria (i)–(iii) for highlights, the extent to which collected datasets (Table 3) reflect them varies greatly, as we will discuss in §5. Table 2 provides examples of sufficient vs. non-sufficient and comprehensive vs. non-comprehensive highlights.

Free-text explanations are free-form textual justifications that are not constrained to the instance inputs. They are thus more expressive and generally more readable than highlights. This makes them useful for explaining reasoning tasks where explanations must contain information outside the given input sentence or document (Camburu et al., 2018; Wiegreffe et al., 2020). They are also called *textual* (e.g., in Kim et al., 2018) or *natural language explanations* (e.g., in Camburu et al., 2018), but these terms have been overloaded to refer to highlights as well (e.g., in Prasad et al., 2020). Other synonyms, *free-form* (e.g., in Mostafazadeh et al., 2020) or *abstractive explanations* (e.g., in Narang et al., 2020) are suitable, but we opt to use "free-text" as it is more descriptive; "free-text" emphasizes that the explanation is textual, regardless of the input modality, just as human explanations are often verbal in many contexts (Lipton, 2018).[4]

[4]We could use the term free-text *rationales* since free-text explanations are introduced to make model explanations more intelligible to people, i.e, they are tailored toward human justifications by design, and typically produced by a model trained with human free-text explanations. The human-computer interaction (HCI) community even advocates to use the term "rationale" to stress that the goal of explainable AI is to produce explanations that sound like humans (Ehsan et al., 2018). However, since our target audience is NLP researchers, we refrain from using "rationale" to avoid confusion with how the term is introduced in Lei et al. (2016) and due to the fact that unsupervised methods for free-text

Dataset	Task	Granularity	Collection	# Instances
MovieReviews (Zaidan et al., 2007)	sentiment classification	none	author	1,800
MovieReviews$_c$ (DeYoung et al., 2020a)	sentiment classification	none	crowd	200‡◇
SST (Socher et al., 2013)	sentiment classification	none	crowd	11,855◇
WikiAttack (Carton et al., 2018)	detecting personal attacks	none	students	1089◇
E-SNLI† (Camburu et al., 2018)	natural language inference	none	crowd	∼569K (1 or 3)
MultiRC (Khashabi et al., 2018)	reading comprehension QA	sentences	crowd	5,825
FEVER (Thorne et al., 2018)	verifying claims from text	sentences	crowd	∼136K‡
HotpotQA (Yang et al., 2018)	reading comprehension QA	sentences	crowd	112,779
Hanselowski et al. (2019)	verifying claims from text	sentences	crowd	6,422 (varies)
NaturalQuestions (Kwiatkowski et al., 2019)	reading comprehension QA	1 paragraph	crowd	n/a‡ (1 or 5)
COS-E v1.0† (Rajani et al., 2019)	commonsense QA	none	crowd	8,560
COS-E v1.11† (Rajani et al., 2019)	commonsense QA	none	crowd	10,962
BoolQ$_c$ (DeYoung et al., 2020a)	reading comprehension QA	none	crowd	199‡◇
EvidenceInference v1.0 (Lehman et al., 2019)	evidence inference	none	experts	10,137
EvidenceInference v1.0$_c$ (DeYoung et al., 2020a)	evidence inference	none	experts	125‡
EvidenceInference v2.0 (DeYoung et al., 2020b)	evidence inference	none	experts	2,503
SciFact (Wadden et al., 2020)	verifying claims from text	1-3 sentences	experts	995‡ (1-3)
Kutlu et al. (2020)	webpage relevance ranking	2-3 sentences	crowd	700 (15)

Table 3: Overview of ExNLP datasets with **highlights**. We do not include the BeerAdvocate dataset (McAuley et al., 2012) because it has been retracted. Values in parentheses indicate number of explanations collected per instance (if greater than 1). DeYoung et al. (2020a) collected (for BoolQ) or recollected annotations for prior datasets (marked with the subscript c). ◇ Collected > 1 explanation per instance but only release 1. † Also contains free-text explanations. ‡ A subset of the original dataset that is annotated. It is not reported what subset of NaturalQuestions has both a long and short answer.

Finally, we use the term **structured explanations** to describe explanations that are not entirely free-form, e.g., there are constraints placed on the explanation-creating process, such as the use of specific inference rules. We discuss the recent emergence of such explanations in §6. Structured explanations do not have one common definition; we elaborate on dataset-specific designs in §4.

4 Overview of Existing Datasets

We overview currently available ExNLP datasets by explanation type: highlights (Table 3), free-text explanations (Table 4), and structured explanations (Table 5). To the best of our knowledge, all existing datasets are in English.

We report the number of instances (input-label pairs) of each dataset, as well as the number of explanations per instance (if > 1). The total number of explanations in a dataset is equal to the # of instances times the value in parentheses.

We report the annotation procedure used to collect each dataset as belonging to the following categories: crowd-annotated ("crowd"); automatically annotated through a web-scrape, database crawl, or merge of existing datasets ("auto"); or annotated by others ("experts", "students", or "authors"). Some datasets use a combination of means; for example, EvidenceInference (Lehman et al., 2019; DeYoung et al., 2020b) was collected and validated by crowdsourcing from doctors. Some authors perform semantic parsing on collected explanations; we classify them by the dataset type before parsing, list the collection type as "crowd + authors", and denote them with $*$. Tables 3-5 elucidate that the dominant collection paradigm ($\geq$90% of entries) is via human (crowd, student, author, or expert) annotation.

4.1 Highlights (Table 3)

The granularity of highlights depends on the task they are collected for. The majority of authors do not place a restriction on granularity, allowing highlights to be made up of words, phrases, or full sentences. Unlike other explanation types, some highlight datasets are re-purposed from datasets for other tasks. For example, in order to collect a question-answering dataset with questions that cannot be answered from a single sentence, the MultiRC authors (Khashabi et al., 2018) ask annotators to specify which sentences are used to obtain the answers (then throw out instances where only one sentence is specified). These collected sentence annotations can be considered sentence-level highlights, as they point to the sentences in a

explanation generation exist (Latcinnik and Berant, 2020).

Dataset	Task	Collection	# Instances
Jansen et al. (2016)	science exam QA	authors	363
Ling et al. (2017)	solving algebraic word problems	auto + crowd	∼101K
Srivastava et al. (2017)*	detecting phishing emails	crowd + authors	7 (30-35)
BABBLELABBLE (Hancock et al., 2018)*	relation extraction	students + authors	200‡‡
E-SNLI (Camburu et al., 2018)	natural language inference	crowd	∼569K (1 or 3)
LIAR-PLUS (Alhindi et al., 2018)	verifying claims from text	auto	12,836
COS-E v1.0 (Rajani et al., 2019)	commonsense QA	crowd	8,560
COS-E v1.11 (Rajani et al., 2019)	commonsense QA	crowd	10,962
WINOWHY (Zhang et al., 2020a)	pronoun coreference resolution	crowd	273 (5)
SBIC (Sap et al., 2020)	social bias inference	crowd	48,923 (1-3)
PUBHEALTH (Kotonya and Toni, 2020b)	verifying claims from text	auto	11,832
Wang et al. (2020)*	relation extraction	crowd + authors	373
Wang et al. (2020)*	sentiment classification	crowd + authors	85
E-δ-NLI (Brahman et al., 2021)	defeasible natural language inference	auto	92,298 (∼8)
BDD-X†† (Kim et al., 2018)	vehicle control for self-driving cars	crowd	∼26K
VQA-E†† (Li et al., 2018)	visual QA	auto	∼270K
VQA-X†† (Park et al., 2018)	visual QA	crowd	28,180 (1 or 3)
ACT-X†† (Park et al., 2018)	activity recognition	crowd	18,030 (3)
Ehsan et al. (2019)††	playing arcade games	crowd	2000
VCR†† (Zellers et al., 2019a)	visual commonsense reasoning	crowd	∼290K
E-SNLI-VE†† (Do et al., 2020)	visual-textual entailment	crowd	11,335 (3)‡
ESPRIT†† (Rajani et al., 2020)	reasoning about qualitative physics	crowd	2441 (2)
VLEP†† (Lei et al., 2020)	future event prediction	auto + crowd	28,726
EMU†† (Da et al., 2020)	reasoning about manipulated images	crowd	48K

Table 4: Overview of EXNLP datasets with **free-text explanations** for textual and visual-textual tasks (marked with †† and placed in the lower part). Values in parentheses indicate number of explanations collected per instance (if greater than 1). ‡ A subset of the original dataset that is annotated. ‡‡ Subset publicly available. ∗ Authors semantically parse the collected explanations.

document that justify each task label (Wang et al., 2019). Another example of re-purposing is the STANFORD SENTIMENT TREEBANK (SST; Socher et al., 2013) which contains crowdsourced sentiment annotations for word phrases extracted from ROTTENTOMATOES movie reviews (Pang and Lee, 2005). Word phrases which have the same sentiment label as the review they belong to can be heuristically merged to produce phrase-level highlights (Carton et al., 2020).

The coursest granularity of highlight in Table 3 is paragraph-level (a paragraph in a larger document is highlighted as evidence). The one dataset with this granularity, (NATURALQUESTIONS), has a paragraph-length long answer paired with a short answer to a question. For instances that have both a short and long answer, the long answer can be repurposed as a highlight to justify the short answer. We exclude datasets that only include an associated document as evidence (and no further annotation of where the explanation is located within the document). The utility of paragraph- or document-retrieval as explanation has not been researched; we leave the exploration of these two courser granularities to future discussion.

4.2 Free-Text Explanations (Table 4)

This is a popular explanation type for both textual and visual-textual tasks, presented in the first half and second half of the table, respectively. Most free-text explanation datasets do not specify a length restriction, but are generally no more than a few sentences per instance. One exception is LIAR-PLUS (Alhindi et al., 2020), which is collected by scraping the conclusion paragraph of human-written fact-checking summaries from POLITIFACT.[5]

4.3 Structured Explanations (Table 5)

Structured explanations take on a variety of dataset-specific forms. One common approach is to construct a chain of facts that detail the multi-hop reasoning to reach an answer ("chains of facts"). Another is to place some constraints on the textual rationales that annotators can write,

[5] https://www.politifact.com/

Dataset	Task	Explanation Type	Collection	# Instances
WorldTree V1 (Jansen et al., 2018)	science exam QA	explanation graphs	authors	1,680
OpenBookQA (Mihaylov et al., 2018)	open-book science QA	1 fact from WorldTree	crowd	5,957
WorldTree V2 (Xie et al., 2020)	science exam QA	explanation graphs	experts	5,100
QED (Lamm et al., 2020)	reading comp. QA	inference rules	authors	8,991
QASC (Khot et al., 2020)	science exam QA	2-fact chain	authors + crowd	9,980
eQASC (Jhamtani and Clark, 2020)	science exam QA	2-fact chain	auto + crowd	9,980 (∼10)
+ Perturbed	science exam QA	2-fact chain	auto + crowd	n/a‡
eOBQA (Jhamtani and Clark, 2020)	open-book science QA	2-fact chain	auto + crowd	n/a‡
Ye et al. (2020)*	SQuAD QA	semi-structured text	crowd + authors	164
Ye et al. (2020)*	NaturalQuestions QA	semi-structured text	crowd + authors	109
R^4C (Inoue et al., 2020)	reading comp. QA	chains of facts	crowd	4,588 (3)
StrategyQA (Geva et al., 2021)	implicit reasoning QA	reasoning steps w/ highlights	crowd	2,780 (3)

Table 5: Overview of ExNLP datasets with **structured explanations** (§6). Values in parentheses indicate number of explanations collected per instance (if greater than 1). ∗ Authors semantically parse the collected explanations. ‡ Subset of instances annotated with explanations is not reported. Total # of explanations is 855 for eQASC Perturbed and 998 for eOBQA.

such as requiring the use or labelling of certain variables in the input ("semi-structured text").

WorldTree V1 (Jansen et al., 2018) proposes explaining elementary-school science questions with a combination of chains of facts and semi-structured text. They term these "explanation graphs". The facts are written by the authors as individual sentences that are centered around one of 62 different relations (e.g., *"affordances"*) or properties (e.g., *"average lifespans of living things"*). Given the chain of facts for an instance (6.3 on average), the authors construct an explanation graph by mapping lexical overlaps (shared words) between the question, answer, and explanation. These explanations are used for the TextGraphs shared task on multi-hop inference (Jansen and Ustalov, 2019), which tested models on their ability to retrieve the appropriate facts from a knowledge base to reconstruct explanation graphs. WorldTree V2 (Xie et al., 2020) expands the dataset using crowdsourcing, and synthesizes the resulting explanation graphs into a set of 344 general inference patterns.

OpenBookQA (Mihaylov et al., 2018) uses single facts from WorldTree v1 to prime annotators to write question-answer pairs by thinking of a second fact that follows. They provide a set of 1,326 of these facts as an "open book" for models to perform the science exam question-answering task. QASC (Khot et al., 2020), uses a similar principle—the authors construct a set of seed facts and a large text corpus of additional science facts. They ask annotators to find a second fact that can be composed with the first, and to create a multiple-choice questions from it. For example, the question "Differential heating of air can be harnessed for what? (Answer: electricity production)" has the following two associated facts: "Differential heating of air produces wind." and "Wind is used for producing electricity." The single fact in OpenBookQA and two facts in QASC can be considered a (partial or full) structured explanation for the resulting instances.

Jhamtani and Clark (2020) extend OBQA and QASC with further two-fact chain explanation annotations. They first automatically extract candidates from a fact corpus, and then crowdsource labels for whether the explanations are valid. The resulting datasets, eQASC and eOBQA, contain multiple (valid and invalid) explanations per instance. The authors also collect perturbed eQASC fact chains that maintain their validity, resulting in another set of fact chains, eQASC-perturbed. While designed for the same task, QASC- and OBQA-based fact chains contain fewer facts than WorldTree explanation graphs on average.

Apart from science exam QA, a number of structured datasets supplement existing QA datasets for reading comprehension. Ye et al. (2020) collect semi-structured explanations for NaturalQuestions (Kwiatkowski et al., 2019) and SQuAD (Rajpurkar et al., 2016). They require annotators to use phrases in both the input question and context, and limit them to a small set of connecting expressions to write sentences explaining how to locate a question's answer within the context (e.g., "X is within 2 words left of Y", "X directly precedes Y", "X is a noun"). Inoue et al. (2020) collect R^4C, fact chain explanations for HotpotQA (Yang et al., 2018). The authors crowdsource semi-structured relational facts based on the sentence-level highlights in HOTPOTQA, but instruct annotators to rewrite them into minimal *subject-verb-object* forms. Lamm et al. (2020) collect explanations for NATURALQUESTIONS that follow a linguistically-motivated form (see an example in Table 1). They identify a single context sentence that entails the answer to the question, mark referential noun-phrase equalities between the question and answer sentence, and extract a semi-structured entailment pattern.

Finally, Geva et al. (2021) propose STRATEGYQA, an implicit reasoning question-answering task and dataset where they ask crowdworkers to decompose complex questions such as "Did Aristotle use a laptop?" into reasoning steps. These decompositions are similar to fact chains, except the steps are questions rather than statements. For each question, crowdworkers provide highlight annotations of an associated passage that contains the answer. In total, the intermediate questions and their highlights explain how to reach an answer.

5 Link Between ExNLP Data, Modeling, and Evaluation Assumptions

All three parts of the machine learning pipeline (data collection, modeling, and evaluation) are inextricably linked. In this section, we discuss what ExNLP modeling and evaluation research reveals about the qualities of available ExNLP datasets, and how best to collect such datasets in the future.

Highlights are usually evaluated following two criteria: (i) *plausibility*, according to humans, how well a highlight supports a predicted label (Yang et al., 2019; DeYoung et al., 2020a),[6] and (ii) *faithfulness* or *fidelity*, how accurately a highlight represents the model's decision process (Alvarez-Melis and Jaakkola, 2018; Wiegreffe and Pinter, 2019). Human-annotated highlights (Table 2) are used to measure the plausiblity of model-produced highlights: the higher the agreement between the two, the more plausible model highlights are.[7] On the other hand, a highlight that is both sufficient (implies the prediction, §3; first example in Table 2) and comprehensive (its complement in the input does *not* imply the prediction, §3; second example in Table 2) is regarded as faithful to the prediction it explains (DeYoung et al., 2020a; Carton et al., 2020). Thus, it might seem that human-annotated highlights are strictly used in the evaluation of plausibility, and not faithfulness. We question this assumption.

Highlight collection is well synthesized in the following instructions for sentiment classification (MOVIEREVIEWS; Zaidan et al., 2007): (i) To justify why a review is positive (negative), highlight the most important words/phrases that tell someone (not) to see the movie, (ii) Highlight enough to provide convincing support, (iii) Do not go out of your way to mark everything. Instructions (ii)–(iii) are written to encourage sufficiency and compactness, but none of the instructions urge comprehensiveness. Indeed, DeYoung et al. (2020a) report that highlights collected in the FEVER, MULTIRC, COS-E, and E-SNLI datasets are comprehensive, but not those in MOVIEREVIEWS and EVIDENCEINFERENCE. They re-annotate subsets of the latter two as well as BOOLQ, which they use to evaluate model comprehensiveness. Conversely, Carton et al. (2020) deem FEVER highlights non-comprehensive by design. Contrary to the characterization of both DeYoung et al. and Carton et al., we observe that the E-SNLI authors collect non-comprehensive highlights, since they instruct annotators to highlight only words in the hypothesis (and not the premise) for neutral pairs, and consider contradiction and neutral explanations correct if only at least one piece of evidence in the input is highlighted (Table 6 in Appendix). Based on these observed discrepancies in characterization, we conclude that post-hoc assessment of comprehensiveness from a general description of data collection is error-prone.

Alternatively, Carton et al. (2020) empirically show that available human-annotated highlights are not necessarily sufficient nor comprehensive for *model predictions*. They demonstrate that human-annotated highlights are insufficient (non-comprehensive) for predictions made by highly accurate models, suggesting that they might be insufficient (non-comprehensive) for gold labels as well. Does this mean that collected highlights in existing datasets are flawed?

Let us first consider insufficiency. Highlighted

[6] Yang et al. (2019) use the term *persuasibility* and define it as "the usefulness degree to human users, serving as the measure of subjective satisfaction or comprehensibility for the corresponding explanation."

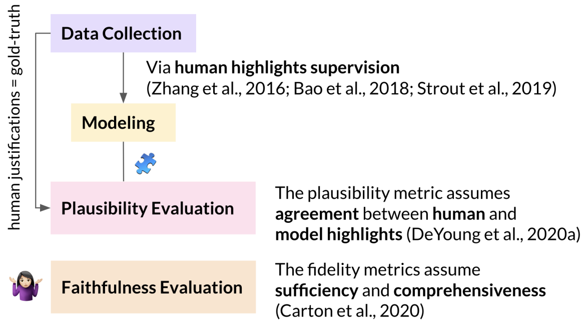
[7] Plausibility is also evaluated with *forward simulation* (Hase and Bansal, 2020): "given an input and an 'explanation,' users must predict what a model would output for the given input". This is a more precise, but more costly measurement as it requires collecting human judgments for every model and its variations.

input elements taken together have to reasonably indicate the label. If they do not, a highlight is an invalid explanation. Consider two datasets whose sufficiency Carton et al. (2020) found to be most concerning: the subset of neutral E-SNLI pairs and the subset of no-attack WIKIATTACK examples. Neutral E-SNLI cases are not justifiable by highlighting because they are obtained only as an intermediate step to collecting free-text explanations, and only free-text explanations truly justify a neutral label (see Table 6 in Appendix; Camburu et al. (2018)). Table 2 shows one E-SNLI highlight that is not sufficient. No-attack WIKIATTACK examples are not explainable by highlighting because the absence of offensive content justifies the no-attack label, and this absence cannot be highlighted. We recommend 1) avoiding human-annotated highlights with low sufficiency when evaluating and collecting highlights, and 2) assessing whether the true label can be explained by highlighting. Does the same hold for non-comprehensiveness?

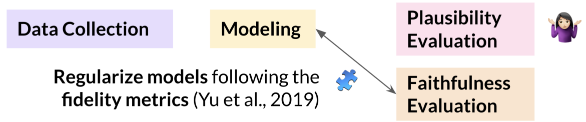
Consider a highlight that is non-comprehensive because it is redundant with its complement in the input (e.g., due to multiple occurrences of the word "great" in a positive movie review, but only some are highlighted). Highlighting only one occurrence of "great" is a valid justification, but guaranteeing and quantifying faithfulness of this highlight is harder because the model might rightfully use the occurrence of "great" that is not highlighted to come to the prediction. Comprehensiveness is adopted to make faithfulness evaluation easier (DeYoung et al., 2020a; Carton et al., 2020).

To recap, in Figure 2, we reconcile assumptions that are made in different stages of development of self-explanatory models that highlight the input elements to justify their decisions. Although our takeaway is that evaluation of both plausibility and faithfulness of the popular highlighting approach (unsupervised models combined with assumptions of sufficiency and comprehensiveness) requires comprehensive human-annotated highlights, we agree with Carton et al. (2020) that:

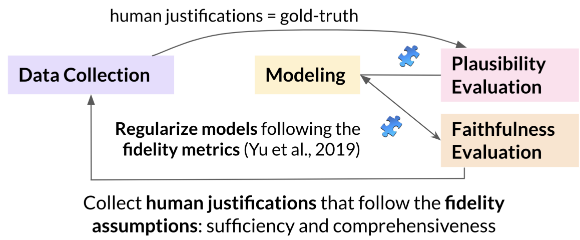
> ...**the idea of onesize-fits-all fidelity benchmarks might be problematic**: human explanations may not be simply treated as gold standard. We need to design careful procedures to collect human explanations, understand properties of the resulting human explanations, and cautiously interpret the evaluation metrics.



(a) Supervised models' development. When we use human highlights as supervision, we assume that they are the gold-truth and that model highlights should match. Thus, comparing human and model highlights for plausibility evaluation is sound. However, with this basic approach we do not introduce any data or modeling properties that help faithfulness evaluation, and that remains a challenge in this setting.



(b) Unsupervised models' development. In §5, we illustrate that comprehensiveness is not a necessary property of human highlights. Non-comprehensiveness, however, hinders evaluating plausibility of model highlights produced in this setting since model and human highlights do not match by design.



(c) Recommended unsupervised models' development. To evaluate both plausibility and faithfulness, we should collect comprehensive human highlights, assuming that they are already sufficient (a necessary property).

Figure 2: Connections between assumptions made in the development of self-explanatory **highlighting** models. The jigsaw icon marks a synergy of modeling and evaluation assumptions. The arrow notes the direction of influence. The text next to the plausibility / faithfulness boxes in the top figure hold for the other figures, but are omitted due to space limits. Cited: DeYoung et al. (2020a); Zhang et al. (2016); Bao et al. (2018); Strout et al. (2019); Carton et al. (2020); Yu et al. (2019).

Mutual influence of data and modeling assumptions also affects free-text explanations. For example, the E-SNLI guidelines (Table 6 in Appendix) have more label-specific and general constraints than the COS-E guidelines (Table 7 in Appendix), such as requiring self-contained explanations. Wiegreffe et al. (2020) show that such data collection decisions can influence modeling assumptions (as in Fig. 2a). This is not an issue per se, but we should be cautious that ExNLP data collection decisions do not popularize explanation properties as *universally necessary* when they are not, e.g., that free-text explanations should be understandable without the original input or that highlights should be comprehensive. We believe this could be avoided with better documentation, e.g., with additions to a standard datasheet (Gebru et al., 2018). While explainability fact sheets have been proposed for models (Sokol and Flach, 2020), they have not been proposed for datasets. For example, an E-SNLI datasheet could note that self-contained explanations were required during data collection, but that this is not a necessary property of a valid free-text explanation. A COS-E datasheet could note that the explanations were not collected to be self-contained. A dataset with comprehensive highlights should emphasize the motivation behind collecting comprehensive highlights: comprehensiveness simplifies evaluation of faithfulness of highlights.

6 Rise of Structured Explanations

The merit of free-text explanations is their expressivity, but expressivity can come at a cost of underspecification and inconsistency due to the difficulty of quality control, as noted by the creators of two popular free-text explanation datasets: E-SNLI and COS-E. In the present section, we highlight and challenge one prior approach to overcoming these difficulties: discarding template-like explanations.

We gather crowdsourcing guidelines for the above-mentioned datasets in Tables 6–7 and compare them. There are two notable similarities between them. First, both asked annotators to first highlight input words and then formulate a free-text explanation from them as a means to control quality. Second, template-like explanations are discarded because they are deemed uninformative. The E-SNLI authors assembled a list of 56 templates (e.g., "There is ⟨hypothesis⟩") to identify explanations whose edit distance to one of the templates was <10. They re-annotate the detected template-like explanations (11% in the entire dataset). The COS-E authors discard sentences "⟨answer⟩ is the only option that is correct/obvious" (the only given example of a template). Another reason why template-like explanations concern researchers is that they can result in artifact-like behaviors in certain modeling architectures. For example, a model which predicts a task output from a generated explanation can produce explanations that are plausible to a human user and give the impression of making label predictions on the basis of this explanation. However, it is possible that the model learns to ignore the semantics of the explanation and instead makes predictions on the basis of the explanation's template type (Kumar and Talukdar, 2020; Jacovi and Goldberg, 2021). In this case, the semantic interpretation of the explanation (which is also that of a human reader) is not faithful (an accurate representation of the model's decision process).

Despite re-annotating, Camburu et al. (2020) report that E-SNLI explanations still largely follow 28 label-specific templates (e.g., an entailment template "X is another form of Y") even after re-annotation. Brahman et al. (2021) report that models trained on gold E-SNLI explanations generate template-like explanations for the δ-NLI task.[8] They identify 11 such templates that make up over 95% of the 200 generated explanations they analyze. These findings lead us to ask: what are the differences between templates considered uninformative and filtered out, and those identified by Camburu et al. (2020); Brahman et al. (2021) that remain after filtering? Are *all* template-like explanations uninformative?

Although prior work seems to indicate that template-like explanations are undesirable, most recently, structured explanations have been intentionally collected (see Table 5). Jhamtani and Clark (2020) define a QA explanation as a chain of steps (typically sentences) that entail an answer, and Lamm et al. (2020) propose a linguistically grounded structure for QA explanations. What these studies share is that they acknowledge structure as *inherent* to explaining the tasks they investigate. Related work (GLUCOSE; Mostafazadeh et al., 2020) takes the matter further, arguing:

> In order to achieve some consensus among explanations and to facilitate further processing and evaluation, **the explanations should not be entirely free-form**.

While we do not survey GLUCOSE in this pa-

[8] Brahman et al. (2021): "the goal of the (discriminative) δ-NLI task is to recognize whether the update weakens or strengthens the entailment of the hypothesis by the premise."

per as it does not contain human-annotated labels (see §2), we expect that findings reported in that work are useful for the explanations considered in this survey. For example, following GLUCOSE, we recommend running pilot studies to explore how people define and generate explanations for a task *before* collecting free-text explanations for that task. If they reveal that informative human explanations are naturally structured, incorporating the structure in the annotation scheme is useful since the structure is natural to explaining the task. This turned out to be the case with natural language inference (Camburu et al., 2020, also reported in Brahman et al., 2021):

> Explanations in E-SNLI largely follow a set of label-specific templates. This is a **natural consequence of the task** and SNLI dataset.

The relevant aspects of GLUCOSE were identified pre-collection following cognitive psychology research. This suggests that there is no all-encompassing definition of explanation (cognitive psychology is not crucial for all tasks) and that researchers could consult domain experts during the pilot studies or follow literature from other fields to define an appropriate explanation in a task-specific manner (for conceptualization of explanations in different fields see Tiddi et al., 2015).

We recommend embracing the structure when possible, but also encourage creators of ExNLP datasets with template-like explanations to stress that template structure can influence downstream modeling decisions. Model developers might consider pruning the explanation template type from intermediate representations of a base model, e.g., with adversarial training (Goodfellow et al., 2015) or more sophisticated methods (Ravfogel et al., 2020), to avoid models learning label-specific templates rather than explanation semantics.

Finally, what if pilot studies do not reveal any obvious structure to human explanations of a task? Then we need to do our best to control the quality of free-text explanations because low dataset quality is a bottleneck to building high-quality models. COS-E is collected with notably less annotation constraints and quality controls than E-SNLI, and has annotation issues that some have deemed make the dataset unusable (Narang et al., 2020, see examples in Table 7). As exemplars of quality control, we gather VCR annotation and collection guidelines in Table 8 (Zellers et al., 2019a) and point the reader to the supplementary material of GLUCOSE (Moon et al., 2020). In §7 and §8, we give more detailed, task-agnostic recommendations for collecting high-quality ExNLP datasets, applicable to all three explanation types.

7 Increasing Explanation Quality

When asked to write free-text sentences from scratch for a table-to-text annotation task (outside of ExNLP), Parikh et al. (2020) note that crowdworkers produce "vanilla targets that lack [linguistic] variety". Lack of variety can result in annotation artifacts, which are prevalent in the popular SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) datasets (Poliak et al., 2018; Gururangan et al., 2018; Tsuchiya, 2018), but also in datasets that contain free-text annotations (Geva et al., 2019). These authors demonstrate the harms of such artifacts: models can overfit to them, leading to both performance over-estimation and problematic generalization behaviors.

Artifacts can also occur from poor-quality annotations and less attentive annotators, both of which are on the rise on crowdsourcing platforms (Chmielewski and Kucker, 2020; Arechar and Rand, 2020). We demonstrate examples of such artifacts in the COS-E v1.11 dataset in Table 7, such as the occurrence of the incorrect explanations "rivers flow trough valleys" on 529 separate instances and "health complications" on 134.[9] To mitigate artifacts, both increased diversity of annotators and quality control are needed. We gather suggestions to increase diversity in §8 and focus on quality control here.

7.1 A Two-Stage Collect-And-Edit Approach

While ad-hoc methods can improve quality (see Tables 6–8 and Moon et al., 2020), an extremely effective and generalizable method is to use a two-stage collection methodology. In ExNLP, Jhamtani and Clark (2020), Zhang et al. (2020a) and Zellers et al. (2019a) apply a two-stage methodology to explanation collection, by either extracting explanation candidates automatically (Jhamtani and Clark, 2020) or having crowdworkers write them (Zhang et al., 2020a; Zellers et al., 2019a). The authors follow collection with an additional step where another batch of crowdworkers assess the quality of the generated explanations (we term this COLLECT-AND-JUDGE). Judging has

[9] Most likely an artifact of annotator(s) copying and pasting the same phrase for multiple answers.

benefits for quality control, and allows the authors to release quality ratings for each instance, which can be used to filter the dataset.

Outside of ExNLP, Parikh et al. (2020) introduce an extended version of this approach (that we term COLLECT-AND-EDIT): they generate a noisy automatically-extracted dataset for the table-to-text generation task, and then ask annotators to edit the datapoints. Bowman et al. (2020) use this approach to collect NLI hypotheses, and find, crucially, that it reduces artifacts in a subset of MNLI instances. They also report that the method introduces more diversity and linguistic variety when compared to datapoints written by annotators from scratch. In XAI, Kutlu et al. (2020) use the approach to collect highlight explanations for Web page ranking. We advocate expanding the COLLECT-AND-JUDGE approach for explanation collection to COLLECT-AND-EDIT. This has potential to improve diversity, reduce individual annotator biases, and perform quality control.

As mentioned above, Jhamtani and Clark (2020) successfully automate the collection of a seed set of explanations, and then use human quality judgements to pare them down. Whether to collect seed explanations automatically or manually depends on task specifics; both have benefits and tradeoffs.[10] However, we will demonstrate through a case study of two multimodal free-text explanation datasets that collecting explanations automatically and *not* including an edit, or at least a judge, phase can lead to artifacts.

Two automatically-collected datasets for visual-textual tasks include explaining visual-textual entailment (E-SNLI-VE; Do et al., 2020) and explaining visual question answering (VQA-E; Li et al., 2018). E-SNLI-VE combines annotations of two datasets: (i) SNLI-VE (Xie et al., 2019), collected by dropping the textual premises of SNLI (Bowman et al., 2015) and replacing them with FLICKR30K images (Young et al., 2014), and (ii) E-SNLI (Camburu et al., 2018), a dataset of crowdsourced explanations for SNLI. This procedure is possible because SNLI's premises are scraped captions from FLICKR30K photos, i.e., every SNLI premise is the caption of a corresponding photo. However, since SNLI's hypotheses were collected from crowdworkers who saw the scraped premises but *not* the original images, the photo replacement process results in a significant number of errors in SNLI-VE (Vu et al., 2018). Do et al. (2020) re-annotate labels and explanations for the neutral pairs in the validation and test sets of SNLI-VE. Others find that the dataset remains low-quality for training models to explain due to artifacts in the entailment class and the neutral class' training set (Marasović et al., 2020). With a full EDIT approach, these artifacts would be significantly reduced, and the resulting dataset could have quality on-par with one collected fully from-scratch.

Similarly, the VQA-E dataset (Li et al., 2018) uses a semantic processing pipeline to automatically convert image captions from the VQA v2.0 dataset (Goyal et al., 2017) into explanations for the question-answer pairs. As a quality measure, Marasović et al. (2020) perform a user study where they ask crowdworkers to judge plausibility of VQA-E explanations, and report a notably lower estimate of plausibility compared to VCR explanations (Zellers et al., 2019a), which are carefully crowdsourced.

Both E-SNLI-VE and VQA-E present novel and cost-effective ways for producing large ExNLP datasets for new tasks. But they also taught us that it is important to consider the quality tradeoffs of automatic collection and potential mitigation strategies such as a selective crowdsourced editing phase. For the tasks presented here, automated collection cannot be fully effective without a full EDIT process. The quality-quantity tradeoff is an exciting direction for future study.

7.2 Teach and Test the Underlying Task

In order to create explanations or judge explanation quality, annotators must understand the underlying task and its label-set well. In most cases, this necessitates teaching and testing annotator understanding of the underlying task as a prerequisite. Exceptions include intuitive tasks that require little effort or expertise such as sentiment classification of short product reviews or commonsense QA.

Tasks based in linguistic concepts may require more training effort. The difficulty of scaling data annotation from expert annotators to lay crowdworkers has been reported for tasks such as semantic role labelling (Roit et al., 2020), complement-coercion (Elazar et al., 2020), and discourse relation annotation (Pyatkin et al., 2020). To increase annotation quality, Roit et al. (2020), Pyatkin et al.

[10] Bansal et al. (2021) collect a few high-quality explanations from LSAT law exam preparation books to accompany the ReClor dataset (Yu et al., 2020), but the authors have to manually edit and shorten the explanations.

(2020), and Mostafazadeh et al. (2020) provide intensive training to crowdworkers, including personal feedback. Even the natural language inference (NLI) task, which has generally been considered straightforward, can require crowdworker training and testing. In a pilot study, the authors of this paper found that $<10\%$ of crowdworker candidates passed a qualifying exam on labelling NLI instances and selecting the correct explanation for a label, despite detailed instructions and examples.

Since label understanding is a prerequisite for explanation collection, task designers should consider relatively inexpensive strategies such as qualification tasks and checker questions to ensure the quality of collected explanations. This need is correlated with the difficulty and domain-specificity of the task, as elaborated above.

7.3 Addressing Ambiguity

We often ask annotators to explain labels assigned by a system or other annotators, i.e., explanations are collected post-hoc. The underlying assumption is that the task has no ambiguity, and the explanation writer will agree with the gold label. However, this assumption has been shown to be inaccurate for some tasks due to a low inter-annotator agreement (Aroyo and Welty, 2013; Pavlick and Kwiatkowski, 2019; Elazar et al., 2020; Nie et al., 2020), and the extent to which it is true likely varies by task, instance, and annotator. If the annotator is uncertain about a label, their explanation for the label may be at best a hypothesis and at worst a guess. HCI research encourages leaving room for ambiguity rather than forcing raters into binary decisions, which can result in poor or inaccurate labels (Sambasivan, 2020).

In order to ensure explanations reflect human decisions as closely as possible, it is ideal to collect both labels and explanations from the same annotators. Given that this is not always possible, including a question to assess whether an explanation annotator agrees with a label, or giving them the (compensated) option to skip an instance are good alternatives. It should be made clear that annotators will not be rejected for selecting these options if done in good faith.

8 Increasing Explanation Diversity

Beyond quality control, increasing annotation diversity is another task-agnostic means to mitigate artifacts and collect more representative data. The COLLECT-AND-EDIT approach (§7.1) is one means to do this; we elaborate on other suggestions from related work (inside and outside EXNLP) here.

8.1 Use a Large Set of Annotators

Outside of EXNLP, Geva et al. (2019) report that recruiting only a small pool of annotators (1 annotator per 100-1000 examples) allows models to overfit on annotator characteristics, causing them to fail to generalize to data collected by other annotators (for reference, E-SNLI reports an average of 860 explanations written per worker).

Both annotators and online content writers can exhibit social biases, which motivates using a large and diverse set of annotators to hopefully dilute biases introduced by individual annotators. Specifically, Davidson et al. (2019) and Sap et al. (2019a) report that crowdworkers are more likely to rate tweets in African-American English as offensive than those written in General American English. Al Kuwatly et al. (2020) find that some demographic attributes can predict annotation differences. Sap et al. (2020) collect biased implications of social media posts and find that 82% of their crowdworkers report their race as white. The conclusion of this set of work is that crowdworkers may not give consistent judgements on social bias annotation tasks because they *exhibit bias themselves*, however subtly manifested, or *represent a skewed population*. This also concerns EXNLP because explanations depend on socio-cultural background (Kopecká and Such, 2020):

> There is a risk that even in the best cases, in which XAI models have been validated with human participants (which is not even always the case in XAI), **the suitability of explanations might not generalise beyond 'WEIRD users:' Western, Educated, Industrialised, Rich and Democratic population** (Bender and Beller, 2013).

We conclude with practical recommendations. By default, annotators are not restricted to a specific number of instances, and competitively-paying requesters will attract a handful of annotators who may individually produce hundreds of annotations. We thus recommend restricting annotators to completing a small number of instances each and/or running multiple studies to amass a more varied pool of worker IDs. Additionally, verifying that no worker has annotated an excessively large portion of the dataset can help mitigate annotator bias, as well as collecting separate training and test sets (if using gold explanations for performance measures) as advocated by Geva et al. (2019). Diversity of the crowd is one mitigation strategy. More elaborate methods have also recently been proposed based on collecting demographic attributes or modeling annotators as a graph (Al Kuwatly et al., 2020; Wich et al., 2020).

8.2 Multiple Annotations Per Instance

HCI research has long considered the ideal of crowdsourcing a single ground-truth as a "myth" that fails to account for the diversity of human thought and experience (Aroyo and Welty, 2015). Similarly, ExNLP researchers should not assume there is always one correct explanation. Many of the assessments crowdworkers are asked to make when writing explanations are subjective in nature. Moreover, there are many different models of explanation based on a user's cognitive biases, social expectations, and socio-cultural background (Miller, 2019; Kopecká and Such, 2020). Prasad et al. (2020) present a theoretical argument to illustrate that there are multiple ways to highlight input words to properly explain an annotated sentiment label. Camburu et al. (2018) collect three free-text explanations for E-SNLI validation and test instances, and find a low inter-annotator BLEU score (Papineni et al., 2002) between them.

If a dataset contains only one explanation, when multiple are plausible, a plausible model explanation can be penalized if it does not agree with the annotated explanation. We expect that modeling multiple explanations could also be a useful learning signal. Some existing datasets contain multiple explanations per instance (see the last column in Tables 3–5). Future ExNLP data collections should do the same if there is sufficient subjectivity in the task or diversity of correct explanations.

However, unlike highlights and structured explanations, free-text explanations do not have straightforward means to measure how similarly multiple explanations support a given label. This makes it difficult to know a good number of explanations to collect for a given task and instance. Practically, we suggest the use of manual inspection and automated metrics such as BLEU to set a value of the number of explanations per instance that adequately captures diversity (i.e., the $\geq n+1^{th}$ explanation is not sufficiently novel or unique given the n collected).

8.3 Get Ahead: Add Contrastive and Negative Explanations

In this section, we overview explanation annotation types beyond what is typically collected that present an exciting direction for future work. Collecting such ExNLP annotations from annotators concurrently with regular explanations will likely be more efficient than collecting them post-hoc.

The broader machine learning community has championed *contrastive explanations* that justify why an event occurred or an action was taken *instead of* some other event or action (Dhurandhar et al., 2018; Miller, 2019; Verma et al., 2020).[11] Most recently, methods have been proposed to produce contrastive explanations for NLP applications (Yang et al., 2020; Wu et al., 2021; Jacovi and Goldberg, 2021). For example, Ross et al. (2020) propose contrastive explanations as the minimal edits of the input that are needed to explain a different (contrast) label for a given input-label pair for a classification task. To the best of our knowledge, there is no dataset that contains contrastive edits with the goal of explainability.[12] However, contrastive edits have been used to assess and improve robustness of NLP models (Kaushik et al., 2020; Gardner et al., 2020; Li et al., 2020) and might used for explainability too. That said, currently these datasets are either small-sized or contain edits only for sentiment classification and/or NLI. Moreover, just as highlights are not sufficiently intelligible for complex tasks, the same might hold for contrastive edits of the input. In this case, we could collect justifications that answer "why...instead of..." in *free-text*.

The effects of edits made on free-text or structured explanations (instead of inputs) for studying ExNLP models' robustness has not yet been thoroughly explored.[13] Related work in computer vision studies *adversarial explanations* that are produced by changing the explanation as much as possible, while keeping the model's prediction mostly unaffected (Alvarez-Melis and Jaakkola, 2018; Zhang et al., 2020b). Producing such explanations for NLP applications is hard due to the discrete inputs, and collecting human edits on explanations might be helpful, such as done in EQASC-

[11] Sometimes called counterfactual explanations.

[12] CoS-E may contain some contrastive explanations: Rajani et al. (2019) note that their annotators often resort to explaining by eliminating the wrong answer choices.

[13] Highlights are made from the input elements, so making edits on a highlight means that edits are made on the input.

PERTURBED (§4.3; Jhamtani and Clark, 2020).

A related annotation paradigm is to collect *negative explanations*, i.e., explanations that are invalid for an input-label pair. Such examples can improve ExNLP models' by providing supervision on what is *not* a correct explanation (Schuff et al., 2020). A human JUDGE or EDIT phase automatically gives negative explanations, which are the low-scoring instances (former) or instances preediting (latter). Jhamtani and Clark (2020); Zhang et al. (2020a) use a JUDGE phase to score explanation candidates, resulting in sets of negative explanations. Zellers et al. (2019a) propose the adversarial matching algorithm to get negative explanations from a set of explanations for other instances.

9 Conclusions

We have presented a review of existing datasets for ExNLP research, highlighted discrepancies in data collection that can have downstream modeling effects, and synthesized the literature both inside and outside of ExNLP into a set of data collection recommendations for future datasets.

We note that a majority of the work reviewed in this paper has originated in the last 1-2 years, indicating an explosion of interest in collecting datasets for ExNLP research. We provide reflections for current and future data collectors in an effort to promote standardization for collected datasets. This paper also serves as a starting resource for newcomers to ExNLP, and, we hope, a starting point for further discussions. We conclude with main takeaways.

Available human-annotated highlights are not necessarily sufficient nor comprehensive (§5).

1. Sufficiency is necessary for highlights, and ExNLP researchers should avoid human-annotated highlights with low sufficiency for evaluating and developing highlights.
2. Comprehensiveness is not necessary for a valid highlight, but it is a means to quantify faithfulness.
3. Non-comprehensive human-annotated highlights cannot be used to automatically evaluate plausibility of highlights that are constrained to be comprehensive. In this case, ExNLP researchers should collect and use comprehensive human-annotated highlights.
4. When deciding which datasets to use, ExNLP researchers should not make post-hoc estimates of comprehensiveness of human-annotated highlights from datasets' general descriptions since that is error-prone.
5. ExNLP researchers should be careful to not popularize their data collection decisions as universally necessary. We advocate for documenting all constraints on collected explanations in a datasheet, highlighting whether each constraint is necessary for explanation to be valid or not, and noting how each constraint might affect modeling and evaluation to the best of the author's knowledge.

Entirely free-text explanations are not always the most natural (§6).

6. ExNLP researchers should study how people define and generate explanations for the task before collecting free-text explanations. If pilot studies show that explanations are naturally structured, embrace the structure.
7. There is no all-encompassing definition of explanation. Thus, ExNLP researchers could consult domain experts or follow literature from other fields to define an appropriate explanation form, and these matters should be open for debate on a given task.

Getting the most out of data collection (§7, §8).

8. Using a COLLECT-AND-EDIT method (§7.1) can improve diversity, reduce individual annotator biases, perform quality control, and potentially reduce dataset artifacts.
9. For quality control, we discuss teaching and testing the underlying task (§7.2) and addressing ambiguity (§7.3).
10. To increase annotation diversity, we discuss the importance of a large set of annotators (§8.1), multiple annotations per instance (§8.2), and collecting explanations that are most useful to the needs of end-users (§8.3).
11. ExNLP challenges discussed in §8, such as the automatic measurement of meaning overlap between two explanations, are similar to fundamental NLP challenges outside of ExNLP. NLP research (especially prior to the emergence of end-to-end black-box neural NLP) might bear on these questions.

Acknowledgements

We are grateful to Yejin Choi, Peter Clark, Gabriel Ilharco, Alon Jacovi, Daniel Khashabi, Mark Riedl, Alexis Ross, and Noah Smith for valuable feedback.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Tariq Alhindi, Smaranda Muresan, and Daniel Preotiuc-Pietro. 2020. Fact vs. opinion: the role of argumentation features in news classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

David Alvarez-Melis and T. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Antonio Alonso Arechar and David Rand. 2020. Turking in the time of covid. PsyArXiv.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Túlio Ribeiro, and Daniel S Weld. 2021. Does the whole exceed its parts? The effect of ai explanations on complementary team performance. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, Brussels, Belgium. Association for Computational Linguistics.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Andrea Bender and Sieghard Beller. 2013. Cognition is... fundamentally cultural. *Behavioral Sciences*, 3:42–54.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*.

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI Workshop on Explainable AI (XAI)*, pages 8–13.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7432–7439.

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, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214, Online. Association for Computational Linguistics.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *the AAAI Conference on Artificial Intelligence*.

Nadia Burkart and Marco F. Huber. 2021. A survey on the explainability of supervised machine learning. *The Journal of Artificial Intelligence Research (JAIR)*, 70.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium. Association for Computational Linguistics.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Michael Chmielewski and Sarah C Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473.

Miruna-Adriana Clinciu and Helen Hastie. 2019. A survey of explainable AI terminology. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 8–13. Association for Computational Linguistics.

Jeff Da, M. Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. 2020. Edited media understanding: Reasoning about implications of manipulated images. arXiv:2012.04726.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020a. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020b. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan

Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 590–601.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-SNLI-VE-2.0: Corrected Visual-Textual Entailment with Natural Language Explanations. In *IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *Proceedings of the Conference of Intelligent User Interfaces (ACM IUI)*.

Yanai Elazar, Victoria Basmov, Shauli Ravfogel, Yoav Goldberg, and Reut Tsarfaty. 2020. The extraordinary failure of complement coercion crowdsourcing. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 106–116, Online. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

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

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*.

Leilani H. Gilpin, David Bau, B. Yuan, A. Bajwa, M. Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the In-*

ternational Conference on Learning Representations (ICLR).

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and D. Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Riccardo Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. 2019. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51:1 – 42.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan. The COLING 2016 Organizing Committee.

Peter Jansen and Dmitry Ustalov. 2019. TextGraphs 2019 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 63–77, Hong Kong. Association for Computational Linguistics.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online. Association for Computational Linguistics.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

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

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Hana Kopecká and Jose M Such. 2020. Explainable AI for Cultural Minds. In *Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction (DEXA HAI) at the 24th European Conference on Artificial Intelligence (ECAI)*.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Mucahid Kutlu, Tyler McDonnell, Matthew Lease, and Tamer Elsayed. 2020. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, 69:143–189.

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*, 7:452–466.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. QED: A framework and dataset for explanations in question answering. *arXiv preprint arXiv:2009.06354*.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. arXiv:2004.05569.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. 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 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784, Online. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. Association for Computational Linguistics.

Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. 2018. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Zachary C Lipton. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57.

Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2810–2829, Online. Association for Computational Linguistics.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*.

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, Brussels, Belgium. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Christoph Molnar. 2019. *Interpretable Machine Learning*. `https://christophm.github.io/interpretable-ml-book/`.

Lori Moon, Lauren Berkowitz, Jennifer Chu-Carroll, and Nasrin Mostafazadeh. 2020. Details of data collection and crowd management for glucose (generalized and contextualized story explanations). *Github*.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019.

Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. arXiv:2004.14546.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Pradeep Dasigi, Dheeru Dua, Matt Gardner, Robert L. Logan IV, Ana Marasović, and Zhen Nie. 2020. Easy, reproducible and quality-controlled data collection with CROWDAQ. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 127–134, Online. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Praveen Paritosh. 2020. Achieving data excellence. In *NeurIPS 2020 Crowd Science Workshop*. Invited talk.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily L. Denton, and A. Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *The ML-Retrospectives, Surveys & Meta-Analyses NeurIPS 2020 Workshop*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. 2020. To what extent do human explanations of model behavior align with actual model behavior? arXiv:2012.13354.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. ESPRIT: Explaining solutions to physical reasoning tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7906–7917, Online. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. 2018. *Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges*. Springer International Publishing, Cham.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Alexis Ross, Ana Marasović, and Matthew E Peters. 2020. Explaining nlp models via minimal contrastive editing (mice). *arXiv preprint arXiv:2012.13985*.

Nithya Sambasivan. 2020. Human-data interaction in ai. In *PAIR Symposium*. Invited talk.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. 2021. " everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019a. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019b. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. 2020. F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1527–1536, Copenhagen, Denmark. Association for Computational Linguistics.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 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)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ilaria Tiddi, M. d'Aquin, and E. Motta. 2015. An ontology design pattern to define explanations. In *Proceedings of the 8th International Conference on Knowledge Capture*.

Masatoshi Tsuchiya. 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 (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sahil Verma, John P. Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. arXiv:2010.10596.

Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. 2018. Grounded textual entailment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2354–2368, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 696–707, Hong Kong, China. Association for Computational Linguistics.

Ziqi Wang, Yujia Qin, Wenxuan Zhou, Jun Yan, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, and Xiang Ren. 2020. Learning from explanations with neural execution tree. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2020. Measuring association between labels and free-text rationales. arXiv:2010.12762.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of* *the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Tongshuang Wu, Marco Túlio Ribeiro, J. Heer, and Daniel S. Weld. 2021. Polyjuice: Automated, general-purpose counterfactual generation. arXiv:2101.00288.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France. European Language Resources Association.

Fan Yang, Mengnan Du, and X. Hu. 2019. Evaluating explanation without ground truth in interpretable machine learning. arXiv:1907.06831.

Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Qinyuan Ye, Xiao Huang, Elizabeth Boschee, and Xiang Ren. 2020. Teaching machine comprehension with compositional explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1599–1615, Online. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. 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 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: a reading comprehension dataset requiring logical reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020a. WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online. Association for Computational Linguistics.

Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020b. Interpretable deep learning under fire. In *29th USENIX Security Symposium (USENIX Security)*.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas. Association for Computational Linguistics.

EXPLAINING NATURAL LANGUAGE INFERENCE (E-SNLI;Camburu et al. 2018)

General Constraints for Quality Control

Guided annotation procedure:

- Step 1: Annotators had to highlight words from the premise/hypothesis that are essential for the given relation.
- Step 2: Annotators had to formulate a free-text explanation using the highlighted words.
- To avoid ungrammatical sentences, only half of the highlighted words had to be used with the same spelling.
- The authors checked that the annotators also used non-highlighted words; correct explanations needs articulate a link between the keywords.
- Annotators had to give self-contained explanations: sentences that make sense without the premise/hypothesis.
- Annotators had to focus on the premise parts that are *not* repeated in the hypothesis (non-obvious elements).
- In-browser check that each explanation contains at least three tokens.
- In-browser check that an explanation is not a copy of the premise or hypothesis.

Label-Specific Constraints for Quality Control

- For entailment, justifications of all the parts of the hypothesis that do not appear in the premise were required.
- For neutral and contradictory pairs, while annotators were encouraged to state all the elements that contribute to the relation, an explanation was considered correct if at least one element is stated.
- For entailment pairs, annotators had to highlight at least one word in the premise.
- For contradiction pairs, annotators had to highlight at least one word in both the premise and the hypothesis.
- For neutral pairs, annotators were allowed to highlight only words in the hypothesis, to strongly emphasize the asymmetry in this relation and to prevent workers from confusing the premise with the hypothesis.

Quality Analysis and Refinement

- The authors graded correctness of 1000 random examples between 0 (incorrect) and 1 (correct), giving partial scores of k/n if only k out of n required arguments were mentioned.
- An explanation was rated as incorrect if it was template-like. The authors assembled a list of 56 templates that they used for identifying explanations (in the entire dataset) whose edit distance to one of the templates was <10. They re-annotated the detected template-like explanations (11% in total).

Post-Hoc Observations

- Total error rate of 9.62%: 19.55% on entailment, 7.26% on neutral, and 9.38% on contradiction.
- In the large majority of the cases, that authors report it is easy to infer label from an explanation.
- Camburu et al. (2020): "Explanations in e-SNLI largely follow a set of label-specific templates. This is a natural consequence of the task and the SNLI dataset and not a requirement in the collection of the e-SNLI. [...] For each label, we created a list of the most used templates that we manually identified among e-SNLI." They collected 28 such templates.

Table 6: Overview of quality control measures and outcomes in E-SNLI.

EXPLAINING COMMONSENSE QA (COS-E; Rajani et al. 2019)
General Constraints for Quality Control Guided annotation procedure: • Step 1: Annotators had to highlight relevant words in the question that justifies the correct answer. • Step 2: Annotators had to provide a brief open-ended explanation based on the highlighted justification that could serve as the commonsense reasoning behind the question. • In-browser check that annotators highlighted at least one relevant word in the question. • In-browser check that an explanation contains at least four words. • In-browser check that an explanation is not a substring of the question or the answer choices without any other extra words.
Label-Specific Constraints for Quality Control (none)
Quality Analysis and Refinement • The authors did unspecified post-collection checks to catch examples that are not caught by their previous filters. • The authors removed template-like explanations, i.e., sentences "⟨answer⟩ is the only option that is correct obvious" (the only provided example of a template).
Post-Hoc Observations • 58% explanations (v1.0) contain the ground truth answer. • The authors report that many explanations remain noisy after quality-control checks, but that they find them to be of sufficient quality for the purposes of their work. • Narang et al. (2020) on v1.11: "Many of the ground-truth explanations for CoS-E are low quality and/or nonsensical (e.g., the question "Little sarah didn't think that anyone should be kissing boys. She thought that boys had what?" with answer "cooties" was annotated with the explanation "american horror comedy film directed"; or the question "What do you fill with ink to print?" with answer "printer" was annotated with the explanation "health complications", etc.)" • Further errors exist (v1.11): The answer "rivers flow trough valleys" appears 529 times, and "health complications" 134 times, signifying copy-paste behavior by some annotators. Uninformative answers such as "this word is the most relevant" (and variants) appear 522 times.

Table 7: Overview of quality control measures and outcomes in COS-E.

Explaining Visual Commonsense Reasoning (VCR; Zellers et al. 2019a)
General Constraints for Quality Control • The authors automatically rate instance "interestingness" and collect annotations for the most "interesting" instances. Multi-stage annotation procedure: • Step 1: Annotators had to write 1-3 questions based on a provided image (at least 4 words each). • Step 2: Annotators had to answer each question (at least 3 words each). • Step 3: Annotators had to provide a rationale for each answer (at least 5 words each). • Annotators had to pass a qualifying exam where they answered some multiple-choice questions and wrote a question, answer, and rationale for a single image. The written responses were verified by the authors. • Authors provided annotators with high-quality question, answer, and rationale examples. • In-browser check that annotators explicitly referred to at least one object detected in the image, on average, in the question, answer, or rationale. • Other in-browser checks related to the question and answer quality. • Every 48 hours, the lead author reviewed work and provided aggregate feedback to make sure the annotators were proving good-quality responses and "structuring rationales in the right way". It is unclear, but assumed, that poor annotators were dropped during these checks.
Label-Specific Constraints for Quality Control (none)
Quality Analysis and Refinement • The authors used a second phase to further refine some HITs. A small group of workers who had done well on the main task were selected to rate a subset of HITs (about 1 in 50), and this process was used to remove annotators with low ratings from the main task.
Post-Hoc Observations • The authors report that humans achieve over 90% accuracy on the multiple-choice rationalization task derived from the dataset. They also report high agreement between the 5 annotators for each instance. These can be indicative of high dataset quality and low noise. • The authors report high diversity—almost every rationale is unique, and the instances cover a range of commonsense categories. • The rationales are long, averaging 16 words in length, another sign of quality. • External validation of quality: Marasović et al. (2020) find that the dataset's explanations are highly plausible with respect to both the image and associated question/answer pairs; they also rarely describe events or objects not present in the image.

Table 8: Overview of quality control measures and outcomes for (the rationale-collection portion) of VCR. The dataset instances (questions and answers) and their rationales were collected simultaneously; we do not include quality controls placed specifically on the question or answer.