Evaluating Models' Local Decision Boundaries via Contrast Sets

Matt Gardner[★◇]   Yoav Artzi[Γ]   Victoria Basmova[◇♣]   Jonathan Berant[◇♠]
Ben Bogin[♠] Sihao Chen[♡] Pradeep Dasigi[◇] Dheeru Dua[□] Yanai Elazar[◇♣]
Ananth Gottumukkala[□] Nitish Gupta[♡] Hanna Hajishirzi[◇△] Gabriel Ilharco[△]
Daniel Khashabi[◇]   Kevin Lin[+]   Jiangming Liu[◇†]   Nelson F. Liu[¶]
Phoebe Mulcaire[△]   Qiang Ning[◇]   Sameer Singh[□]   Noah A. Smith[◇△]
Sanjay Subramanian[◇] Reut Tsarfaty[◇♣] Eric Wallace[+] Ally Zhang[Γ] Ben Zhou[♡]
[◇]Allen Institute for AI   [Γ]Cornell University   [♣]Bar-Ilan University
[♠]Tel-Aviv University   [♡]University of Pennsylvania   [△]University of Washington
[□]UC Irvine   [+]UC Berkeley   [†]University of Edinburgh   [¶]Stanford University
mattg@allenai.org

Abstract

Standard test sets for supervised learning evaluate in-distribution generalization. Unfortunately, when a dataset has systematic gaps (e.g., annotation artifacts), these evaluations are misleading: a model can learn simple decision rules that perform well on the test set but do not capture the abilities a dataset is intended to test. We propose a more rigorous annotation paradigm for NLP that helps to close systematic gaps in the test data. In particular, after a dataset is constructed, we recommend that the dataset authors manually perturb the test instances in small but meaningful ways that (typically) change the gold label, creating *contrast sets*. Contrast sets provide a local view of a model's decision boundary, which can be used to more accurately evaluate a model's true linguistic capabilities. We demonstrate the efficacy of contrast sets by creating them for 10 diverse NLP datasets (e.g., DROP reading comprehension, UD parsing, and IMDb sentiment analysis). Although our contrast sets are not explicitly adversarial, model performance is significantly lower on them than on the original test sets—up to 25% in some cases. We release our contrast sets as new evaluation benchmarks and encourage future dataset construction efforts to follow similar annotation processes.

1 Introduction

Progress in natural language processing (NLP) has long been measured with standard benchmark datasets (e.g., Marcus et al., 1993). These benchmarks help to provide a uniform evaluation of new modeling developments. However, recent work shows a problem with this standard evaluation paradigm based on *i.i.d.* test sets: datasets often

Original Example:

Two similarly-colored and similarly-posed chow dogs are face to face in one image.

Example Textual Perturbations:

Two similarly-colored and similarly-posed **cats** are face to face in one image.
Three similarly-colored and similarly-posed chow dogs are face to face in one image.
Two **differently-colored but** similarly-posed chow dogs are face to face in one image.

Example Image Perturbation:

Two similarly-colored and similarly-posed chow dogs are face to face in one image.

Figure 1: An example contrast set for NLVR2 (Suhr and Artzi, 2019). The label for the original example is TRUE and the label for all of the perturbed examples is FALSE. The contrast set allows probing of a model's decision boundary local to examples in the test set, which better evaluates whether the model has captured the relevant phenomena than standard metrics on *i.i.d.* test data.

have systematic gaps (such as those due to various kinds of annotator bias) that (unintentionally) allow simple decision rules to perform well on test data (Chen et al., 2016; Gururangan et al., 2018; Geva et al., 2019). This is strikingly evident when models achieve high test accuracy but fail on simple input perturbations (Jia and Liang, 2017; Feng et al., 2018; Ribeiro et al., 2018a), challenge examples (Naik et al., 2018), and covariate and label shifts (Ben-David et al., 2010; Shimodaira, 2000; Lipton et al., 2018).

To more accurately evaluate a model's true capabilities on some task, we must collect data that

---

★ Matt Gardner led the project. All other authors are listed in alphabetical order.

fills in these systematic gaps in the test set. To accomplish this, we expand on long-standing ideas of constructing minimally-constrastive examples (e.g. Levesque et al., 2011). We propose that dataset authors manually perturb instances from their test set, creating *contrast sets* which characterize the correct decision boundary near the test instances (Section 2). Following the dataset construction process, one should make small but (typically) label-changing modifications to the existing test instances (e.g., Figure 1). These perturbations should be small, so that they preserve whatever lexical/syntactic artifacts are present in the original example, but change the true label. They should be created *without* a model in the loop, so as not to bias the contrast sets towards quirks of particular models. Having a set of contrasting perturbations for test instances allows for a *consistency* metric that measures how well a model's decision boundary aligns with the "correct" decision boundary around each test instance.

Perturbed test sets only need to be large enough to draw substantiated conclusions about model behavior and thus do not require undue labor on the original dataset authors. We show that using about a person-week of work can yield high-quality perturbed test sets of approximately 1000 instances for many commonly studied NLP benchmarks, though the amount of work varies greatly (Section 3).

We apply this annotation paradigm to a diverse set of 10 existing NLP datasets—including visual reasoning, reading comprehension, sentiment analysis, and syntactic parsing—to demonstrate its wide applicability and efficacy (Section 4). Although contrast sets are not intentionally adversarial, state-of-the-art models perform dramatically worse on our contrast sets than on the original test sets, especially when evaluating consistency. We believe that contrast sets provide a more accurate reflection of a model's true performance, and we release our datasets as new benchmarks.[1] We recommend that creating contrast sets become standard practice for NLP datasets.
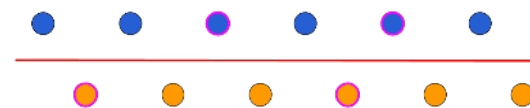
2 Contrast Sets

2.1 The Problem

We first give a sketch of the problem that contrast sets attempt to solve in a toy two-dimensional classification setting as shown in Figure 2. Here, the
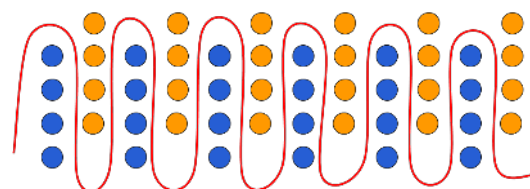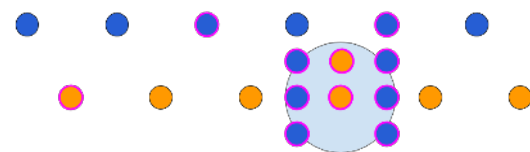
(a) A two-dimensional dataset that requires a complex decision boundary to achieve high accuracy.

(b) If the same data distribution is instead sampled with systematic gaps (e.g., due to annotator bias), a simple decision boundary *can perform well on i.i.d. test data* (shown outlined in pink).

(c) Since filling in all gaps in the distribution is infeasible, a *contrast set* instead fills in a local ball around a test instance to evaluate the model's decision boundary.

Figure 2: An illustration of how contrast sets provide a more comprehensive model evaluation when datasets have systematic gaps.

true underlying data distribution requires a complex decision boundary (Figure 2a). However, as is common in practice, our toy dataset is rife with systematic gaps (e.g., due to annotator bias, repeated patterns, etc.). This causes simple decision boundaries to emerge (Figure 2b). And, because our biased dataset is split *i.i.d.* into train and test sets, this simple decision boundary will perform well on test data. Ideally, we would like to fill in all of a dataset's systematic gaps, however, this is usually impossible. Instead, we create a *contrast set*: a collection of instances tightly clustered in input space around a single test instance, or *pivot* (Figure 2c; an ϵ-ball in our toy example). This contrast set allows us to measure how well a model's decision boundary aligns with the correct decision boundary local to the pivot. In this case, the contrast set demonstrates that the model's simple decision boundary is incorrect. We repeat this process around numerous pivots to form entire evaluation datasets.

When we move from toy settings to complex NLP tasks, the precise nature of a "systematic gap" in the data becomes harder to define. Indeed, the geometric view in our toy examples does not correspond directly to experts' perception of data; there are many ways to "locally perturb" natural lan-

Dataset	Original Instance	Contrastive Instance (color = edit)
IMDb	Hardly one to be faulted for his ambition or his vision, it is genuinely unexpected, then, to see all Park's effort add up to so very little. . . . The premise is promising, gags are copious and offbeat humour abounds but it all fails miserably to create any meaningful connection with the audience. *(Label: Negative)*	Hardly one to be faulted for his ambition or his vision, **here we see all Park's effort come to fruition.** . . . The premise is **perfect**, gags are **hilarious** and offbeat humour abounds, **and it creates a deep** connection with the audience. *(Label: Positive)*
MATRES	Colonel Collins followed a normal progression once she was picked as a NASA astronaut. *("picked" was before "followed")*	Colonel Collins followed a normal progression **before** she was picked as a NASA astronaut. *("picked" was after "followed")*
UD English	They demanded talks with local US commanders. I attach a paper on gas storage value modeling. I need to get a job at the earliest opportunity.	They demanded talks with **great urgency**. I attach a paper on **my own initiative**. I need to get a job at **House of Pies**.
PERSPECTRUM	**Claim:** Should uniforms be worn at school. **Perspective:** School uniforms emphasize the socio-economic divisions they are supposed to eliminate. **Label:** Against	**Claim:** Should uniforms be **banned** at school. **Perspective:** School uniforms emphasize the socio-economic divisions they are supposed to eliminate. **Label:** For
DROP	**Context:** In the spring of 1625 the Spanish regained Bahia in Brazil and Breda in the Netherlands from the Dutch. In the autumn they repulsed the English at Cadiz. **Question:** What event happened first, the Spanish repulsed the English at Cadiz or the Spanish regained Bahia?	**Context:** In the spring of 1625 the Spanish regained Bahia in Brazil and Breda in the Netherlands from the Dutch. In **winter the year earlier** they had repulsed the English at Cadiz. **Question:** What event happened first, the Spanish repulsed the English at Cadiz or the Spanish regained Bahia?
QUOREF	**Context:** Matt Helm is a secret agent. His assignment is to stop the sinister Tung-Tze, armed with spy gadgets. Helm prevails with Gail by his side as he destroys Tung-Tze. **Question:** Who is armed with spy gadgets?	**Context:** Matt Helm is a secret agent. His assignment is to stop the sinister Tung-Tze, **even though he is** armed with spy gadgets. Helm prevails with Gail by his side as he destroys Tung-Tze. **Question:** Who is armed with spy gadgets?
MC-TACO	**Context:** She renews in Ranchipur an acquaintance with a former lover, Tom Ransome, now a dissolute alcoholic. **Question:** How frequently does Tom drink? **Candidate Answer:** Every other night **Label:** Likely	**Context:** She renews in Ranchipur an acquaintance with a former lover, Tom Ransome, **who keeps very healthy habits**. **Question:** How frequently does Tom drink? **Candidate Answer:** Every other night **Label:** Unlikely

Table 1: We create contrast sets for 10 datasets and show instances from seven of them here.

guage. We do not expect intuition, even of experts, to exhaustively reveal gaps.

Nevertheless, the presence of these gaps is well-documented (Gururangan et al., 2018; Poliak et al., 2018; Min et al., 2019), and Niven and Kao (2019) give an initial attempt at formally characterizing them. In particular, one common source is annotator bias from data collection processes (Geva et al., 2019). For example, in the SNLI dataset (Bowman et al., 2015), Gururangan et al. (2018) show that the words *sleeping*, *tv*, and *cat* almost never appear in an entailment example, either in the training set or the test set, though they often appear in contradiction examples. This is not because these words are particularly important to the phenomenon of entailment; their absence in entailment examples is a *systematic gap* in the data that can be exploited by models to achieve artificially high test accuracy.

This is but one kind of systematic gap; there are also biases due to the writing styles of small groups of annotators (Geva et al., 2019), the distributional biases in the data that was chosen for annotation, as well as numerous other biases that are more subtle and harder to discern (Shah et al., 2020).

Completely removing these gaps in the initial data collection process would be ideal, but is likely impossible—language has too much inherent variability in a very high-dimensional space. Instead, we use contrast sets to fill in gaps in the test data to give more thorough evaluations than what the original data provides.

2.2 Definitions

We begin by defining a ***decision boundary*** as a partition of some space into labels.[2] This partition can be represented by the set of all points in the space with their associated labels: $\{(x, y)\}$. This definition differs somewhat from the canonical definition, which is a collection of hypersurfaces that separate labels. There is a bijection between partitions and these sets of hypersurfaces in continuous spaces, however, so they are equivalent definitions. We choose to use the partition to represent the decision boundary as it makes it very easy to define a *local* decision boundary and to generalize the notion to discrete spaces, which we deal with in NLP.

A ***local decision boundary*** around some ***pivot*** x is the set of all points x' and their associated labels y' that are within some distance ϵ of x. That is, a local decision boundary around x is the set $\{(x', y') \mid d(x, x') < \epsilon\}$. Note here that even though a "boundary" or "surface" is hard to visualize in a discrete input space, using this partition representation instead of hypersurfaces gives us a uniform definition of a local decision boundary in any input space; all that is needed is a distance function d.

A ***contrast set*** $C(x)$ is any sample of points from a local decision boundary around x. In other words, $C(x)$ consists of inputs x' that are similar to x according to some distance function d. Typically these points are sampled such that $y' \neq y$. To evaluate a model using these contrast sets, we define the ***contrast consistency*** of a model to be whether it makes correct predictions $\hat{y}$ on every element in the set: $\text{all}(\{\hat{y} = y' \; \forall (x', y') \in C(x)\})$. Since the points x' were chosen from the local decision boundary, we expect contrast consistency on expert-built contrast sets to be a significantly more accurate evaluation of whether model predictions match the task definition than a random selection of input / output pairs.

2.3 Contrast sets in practice

Given these definitions, we now turn to the actual construction of contrast sets in practical NLP settings. There were two things left unspecified in the definitions above: the distance function d to use in discrete input spaces, and the method for sampling from a local decision boundary. While there has been some work trying to formally characterize dis-

---

[2]In this discussion we are talking about the *true* decision boundary, not a *model's* decision boundary.

tances for adversarial robustness in NLP (Michel et al., 2019; Jia et al., 2019), we find it more useful in our setting to simply rely on expert judgments to generate a similar but meaningfully different x' given x, addressing both the distance function and the sampling method.

Future work could try to give formal treatments of these issues, but we believe expert judgments are sufficient to make initial progress in improving our evaluation methodologies. And while expert-crafted contrast sets can only give us an upper bound on a model's local alignment with the true decision boundary, an upper bound on local alignment is often more informative than a potentially biased *i.i.d.* evaluation that permits artificially simple decision boundaries. To give a tighter upper bound, we draw pivots x from some *i.i.d.* test set, and we do not provide *i.i.d.* contrast sets at training time, which could provide additional artificially simple decision boundaries to a model.

Figure 1 displays an example contrast set for the NLVR2 visual reasoning dataset (Suhr and Artzi, 2019). Here, both the sentence and the image are modified in small ways (e.g., by changing a word in the sentence or finding a similar but different image) to make the output label change.

A contrast set is *not* a collection of adversarial examples (Szegedy et al., 2014). Adversarial examples are almost the methodological opposite of contrast sets: they change the input such that a model's decision changes but the gold label does not (Jia and Liang, 2017; Wallace et al., 2019a). On the other hand, contrast sets are model-agnostic, constructed by experts to characterize whether a model's decision boundary locally aligns to the true decision boundary around some point. Doing this requires input changes that also induce changes to the gold label.

We recommend that the original dataset authors— the experts on the linguistic phenomena intended to be reflected in their dataset—construct the contrast sets. This is best done by first identifying a list of phenomena that characterize their dataset. In syntactic parsing, for example, this list might include prepositional phrase attachment ambiguities, coordination scope, clausal attachment, etc. After the standard dataset collection process, the authors should sample pivots from their test set and perturb them according to the listed phenomena.

2.4 Design Choices of Contrast Sets

Here, we discuss possible alternatives to our approach for constructing contrast sets and our reasons for choosing the process we did.

Post-hoc Construction of Contrast Sets Improving the evaluation for existing datasets well after their release is usually too late: new models have been designed, research papers have been published, and the community has absorbed potentially incorrect insights. Furthermore, post-hoc contrast sets may be biased by existing models. We instead recommend that new datasets include contrast sets upon release, so that the authors can characterize beforehand when they will be satisfied that a model has acquired the dataset's intended capabilities. Nevertheless, contrast sets constructed post-hoc are still better than typical *i.i.d.* test sets, and where feasible we recommend creating contrast sets for existing datasets (as we do in this work).

Crowdsourcing Contrast Sets We recommend that the dataset authors construct contrast sets themselves rather than using crowd workers. The original authors are the ones who best understand their dataset's intended phenomena and the distinction between in-distribution and out-of-distribution examples—these ideas can be difficult to distill to non-expert crowd workers. Moreover, the effort to create contrast sets is a small fraction of the effort required to produce a new dataset in the first place.

Automatic Construction of Contrast Sets Automatic perturbations, such as paraphrasing with back-translation or applying word replacement rules, can fill in some parts of the gaps around a pivot (e.g., Ribeiro et al., 2018b, 2019). However, it is very challenging to come up with rules or other automated methods for pushing pivots *across a decision boundary*—in most cases this presupposes a model that can already perform the intended task. We recommend annotators spend their time constructing these types of examples; easier examples can be automated.

Adversarial Construction of Contrast Sets Some recent datasets are constructed using baseline models in the data collection process, either to filter out examples that existing models answer correctly (e.g., Dua et al., 2019; Dasigi et al., 2019) or to generate adversarial inputs (e.g., Zellers et al., 2018, 2019; Wallace et al., 2019b; Nie et al., 2019).

Unlike this line of work, we choose *not* to have a model in the loop because this can bias the data to the failures of a particular model (cf. Zellers et al., 2019), rather than generally characterizing the local decision boundary. We do think it is acceptable to use a model on a handful of initial perturbations to understand which phenomena are worth spending time on, but this should be separate from the actual annotation process—observing model outputs while perturbing data creates subtle, undesirable biases towards the idiosyncrasies of that model.

2.5 Limitations of Contrast Sets

Solely Negative Predictive Power Contrast sets only have negative predictive power: they reveal if a model *does not* align with the correct local decision boundary but cannot confirm that a model *does* align with it. This is because annotators cannot exhaustively label all inputs near a pivot and thus a contrast set will necessarily be incomplete. However, note that this problem is not unique to contrast sets—similar issues hold for the original test set as well as adversarial test sets (Jia and Liang, 2017), challenge sets (Naik et al., 2018), and input perturbations (Ribeiro et al., 2018a; Feng et al., 2018). See Feng et al. (2019) for a detailed discussion of how dataset analysis methods only have negative predictive power.

Dataset-Specific Instantiations The process for creating contrast sets is *dataset-specific*: although we present general guidelines that hold across many tasks, experts must still characterize the type of phenomena each individual dataset is intended to capture. Fortunately, the original dataset authors should *already* have thought deeply about such phenomena. Hence, creating contrast sets should be well-defined and relatively straightforward.

3 How to Create Contrast Sets

Here, we walk through our process for creating contrast sets for three datasets. Examples are shown in Figure 1 and Table 1.

DROP DROP (Dua et al., 2019) is a reading comprehension dataset that is intended to cover compositional reasoning over numbers in a paragraph, including filtering, sorting, and counting sets, and doing numerical arithmetic. The data has three main sources of paragraphs, all from Wikipedia articles: descriptions of American football games, descriptions of census results, and summaries of

wars. There are many common patterns used by the crowd workers that make some questions artificially easy: 2 is the most frequent answer to *How many...?* questions, questions asking about the ordering of events typically follow the linear order of the paragraph, and a large fraction of the questions do not require compositional reasoning.

Our strategy for constructing contrast sets for DROP was three-fold. First, we added more compositional reasoning steps. The questions about American football passages in the original data very often had multiple reasoning steps (e.g., *How many yards difference was there between the Broncos' first touchdown and their last?*), but the questions about the other passage types did not. We drew from common patterns in the training data and added additional reasoning steps to questions in our contrast sets. Second, we inverted the semantics of various parts of the question. This includes perturbations such as changing *shortest* to *longest*, *later* to *earlier*, as well as changing questions asking for counts to questions asking for sets (*How many countries...* to *Which countries...*). Finally, we changed the ordering of events. A large number of questions about war paragraphs ask which of two events happened first. We changed (1) the order the events were asked about in the question, (2) the order that the events showed up in the passage, and (3) the dates associated with each event to swap their temporal order.

NLVR2 We next consider NLVR2, a dataset where a model is given a sentence about two provided images and must determine whether the sentence is true (Suhr et al., 2019). The data collection process encouraged highly compositional language, which was intended to require understanding the relationships between objects, properties of objects, and counting. We constructed NLVR2 contrast sets by modifying the sentence or replacing one of the images with freely-licensed images from web searches. For example, we might change *The left image contains twice the number of dogs as the right image* to *The left image contains **three times** the number of dogs as the right image*. Similarly, given an image pair with four dogs in the left and two dogs in the right, we can replace individual images with photos of variably-sized groups of dogs. The textual perturbations were often changes in quantifiers (e.g., *at least one* to *exactly one*), entities (e.g., *dogs* to *cats*), or properties thereof (e.g.,

orange glass to *green glass*). An example contrast set for NLVR2 is shown in Figure 1.

UD Parsing Finally, we discuss dependency parsing in the universal dependencies (UD) formalism (Nivre et al., 2016). We look at dependency parsing to show that contrast sets apply not only to modern "high-level" NLP tasks but also to long-standing linguistic analysis tasks. We first chose a specific type of attachment ambiguity to target: the classic problem of prepositional phrase (PP) attachment (Collins and Brooks, 1995), e.g. *We ate spaghetti with forks* versus *We ate spaghetti with meatballs*. We use a subset of the English UD treebanks: GUM (Zeldes, 2017), the English portion of LinES (Ahrenberg, 2007), the English portion of ParTUT (Sanguinetti and Bosco, 2015), and the dependency-annotated English Web Treebank (Silveira et al., 2014). We searched these treebanks for sentences that include a potentially structurally ambiguous attachment from the head of a PP to either a noun or a verb. We then perturbed these sentences by altering one of their noun phrases such that the semantics of the perturbed sentence required a different attachment for the PP. We then re-annotated these perturbed sentences to indicate the new attachment(s).

Summary While the overall process we recommend for constructing contrast sets is simple and unified, its actual instantiation varies for each dataset. Dataset authors should use their best judgment to select which phenomena they are most interested in studying and craft their contrast sets to explicitly test those phenomena. Care should be taken during contrast set construction to ensure that the phenomena present in contrast sets are similar to those present in the original test set; the purpose of a contrast set is not to introduce new challenges, but to more thoroughly evaluate the original intent of the test set.

4 Datasets and Experiments

4.1 Original Datasets

We create contrast sets for 10 NLP datasets (full descriptions are provided in Section A):
- **NLVR2** (Suhr et al., 2019)
- **IMDb sentiment analysis** (Maas et al., 2011)
- **MATRES Temporal RE** (Ning et al., 2018)
- **English UD parsing** (Nivre et al., 2016)
- **PERSPECTRUM** (Chen et al., 2019)
- **DROP** (Dua et al., 2019)

1312

Dataset	# Examples	# Sets	Model	Original Test	Contrast		Consistency
NLVR2	994	479	LXMERT	76.4	61.1	(–15.3)	30.1
IMDb	488	488	BERT	93.8	84.2	(–9.6)	77.8
MATRES	401	239	CogCompTime2.0	73.2	63.3	(–9.9)	40.6
UD English	150	150	Biaffine + ELMo	64.7	46.0	(–18.7)	17.3
PERSPECTRUM	217	217	RoBERTa	90.3	85.7	(–4.6)	78.8
DROP	947	623	MTMSN	79.9	54.2	(–25.7)	39.0
QUOREF	700	415	XLNet-QA	70.5	55.4	(–15.1)	29.9
ROPES	974	974	RoBERTa	47.7	32.5	(–15.2)	17.6
BoolQ	339	70	RoBERTa	86.1	71.1	(–15.0)	59.0
MC-TACO	646	646	RoBERTa	38.0	14.0	(–24.0)	8.0

Table 2: Models struggle on the contrast sets compared to the original test sets. For each dataset, we use a (sometimes near) state-of-the-art model and evaluate it on the "# Examples" examples in the contrast sets (*not* including the original example). We report percentage accuracy for NLVR2, IMDb, PERSPECTRUM, MATRES, and BoolQ; F_1 scores for DROP and QUOREF; Exact Match (EM) scores for ROPES and MC-TACO; and unlabeled attachment score on modified attachments for the UD English dataset. We also report *contrast consistency*: the percentage of the "# Sets" contrast sets for which a model's predictions are correct for all examples in the set (*including* the original example). More details on datasets, models, and metrics can be found in §A and §B.

- **Quoref** (Dasigi et al., 2019)
- **ROPES** (Lin et al., 2019)
- **BoolQ** (Clark et al., 2019)
- **MC-TACO** (Zhou et al., 2019)

We choose these datasets because they span a variety of tasks (e.g., reading comprehension, sentiment analysis, visual reasoning) and input-output formats (e.g., classification, span extraction, structured prediction). We include high-level tasks for which dataset artifacts are known to be prevalent, as well as longstanding formalism-based tasks, where data artifacts have been less of an issue (or at least have been less well-studied).

4.2 Contrast Set Construction

The contrast sets were constructed by NLP researchers who were deeply familiar with the phenomena underlying the annotated dataset; in most cases, these were the original dataset authors. Our contrast sets consist of up to about 1,000 total examples and average 1–5 examples per contrast set (Table 2). We show representative examples from the different contrast sets in Table 1. For most datasets, the average time to perturb each example was 1–3 minutes, which translates to approximately 17–50 hours of work to create 1,000 examples. However, some datasets, particularly those with complex output structures, took substantially

longer: each example for dependency parsing took an average of 15 minutes (see Appendix B for more details).

4.3 Models Struggle on Contrast Sets

For each dataset, we use a model that is at or near state-of-the-art performance. Most models involve fine-tuning a pretrained language model (e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), ROBERTA (Liu et al., 2019), XLNet (Yang et al., 2019), etc.) or applying a task-specific architecture on top of one (e.g., Hu et al. (2019) add a DROP-specific model on top of BERT). We train each model on the original training set and evaluate it on both the original test set and our contrast sets.

Existing models struggle on the contrast sets (Table 2), particularly when evaluating contrast consistency. Model performance degrades differently across datasets; however, note that these numbers are not directly comparable due to differences in dataset size, model architecture, contrast set design, etc. On IMDb and PERSPECTRUM, the model achieves a reasonably high consistency, suggesting that, while there is definitely still room for improvement, the phenomena targeted by those datasets are already relatively well captured by existing models.

Of particular note is the very low consistency score for dependency parsing. The parser that we

use achieves 95.7% unlabeled attachment score on the English Penn Treebank (Dozat and Manning, 2017, trained with ELMo embeddings). A consistency score of 17.3 on a common attachment ambiguity suggests that this parser may not be as strong as common evaluations lead us to believe. Overall, our results suggest that models have "overfit" to artifacts that are present in existing datasets; they achieve high test scores but do not completely capture a dataset's intended phenomena.

4.4 Humans Succeed On Contrast Sets

An alternative explanation for why models fail on the contrast sets is that they are simply harder or noisier than regular test sets, i.e., humans would also perform worse on the contrast sets. We show that this is not the case. For four datasets, we choose at least 100 test instances and one corresponding contrast set instance (i.e., an example before and after perturbation). We (the authors) test ourselves on these examples (ensuring that those who were tested were different from those who created the examples). Human performance is comparable across the original test and contrasts set examples on these datasets (Table 3).

Dataset	Original Test	Contrast Set	
IMDb	94.3	93.9	(–0.4)
PERSPECTRUM	91.5	90.3	(–1.2)
QUOREF	95.2	88.4	(–6.8)
ROPES	76.0	73.0	(–3.0)

Table 3: Humans achieve similar performance on the contrast sets and the original test sets. The metrics here are the same as those in Table 2.

4.5 Fine-Grained Analysis of Contrast Sets

Each example in the contrast sets can be labeled according to which particular phenomenon it targets. This allows automated error reporting. For example, for the MATRES dataset we tracked whether a perturbation changed appearance order, tense, or temporal conjunction words. These fine-grained labels show that the model does comparatively better at modeling appearance order (66.5% of perturbed examples correct) than temporal conjunction words (60.0% correct); see Appendix B.3 for full details. A similar analysis on DROP shows that MTMSN does substantially worse on event re-ordering (47.3 F_1) than on adding compositional reasoning steps (67.5 F_1). We recommend authors categorize their

perturbations up front in order to simplify future analyses and bypass some of the pitfalls of post-hoc error categorization (Wu et al., 2019).

Additionally, it's worth discussing the dependency parsing result. The attachment decision that we targeted was between a verb, a noun, and a preposition. With just two reasonable attachment choices, a contrast consistency of 17.3 means that the model is almost always unable to change its attachment based on the content of the prepositional phrase. Essentially, in a trigram such as *demanded talks with* (Table 1), the model has a bias for whether *demanded* or *talks* has a stronger affinity to *with*, and makes a prediction accordingly. Given that trigrams are rare and annotating parse trees is expensive, it is not clear that traditional evaluation metrics with *i.i.d* test sets would ever find this problem. By robustly characterizing local decision boundaries, contrast sets surface errors that are very challenging to find with other means.

5 Related Work

The fundamental idea of finding or creating data that is "minimally different" has a very long history. In linguistics, for instance, the term *minimal pair* is used to denote two words with different meaning that differ by a single sound change, thus demonstrating that the sound change is phonemic in that language (Pike, 1946). Many people have used this idea in NLP (see below), creating challenge sets or providing training data that is "minimally different" in some sense, and we continue this tradition. Our main contribution to this line of work, in addition to the resources that we have created, is giving a simple and intuitive geometric interpretation of "bias" in dataset collection, and showing that this long-standing idea of minimal data changes can be effectively used to solve this problem on a wide variety of NLP tasks. We additionally generalize the idea of a minimal *pair* to a *set*, and use a *consistency* metric, which we contend more closely aligns with what NLP researchers mean by "language understanding".

Training on Perturbed Examples Many previous works have provided minimally contrastive examples on which to train models. Selsam et al. (2019), Tafjord et al. (2019), Lin et al. (2019), and Khashabi et al. (2020) designed their data collection process to include contrastive examples. Data augmentation methods have also been used to mitigate gender (Zhao et al., 2018), racial (Dixon et al.,

2018), and other biases (Kaushik et al., 2020) during training, or to introduce useful inductive biases (Andreas, 2020).

Challenge Sets The idea of creating challenging contrastive evaluation sets has a long history (Levesque et al., 2011; Ettinger et al., 2017; Glockner et al., 2018; Naik et al., 2018; Isabelle et al., 2017). Challenge sets exist for various phenomena, including ones with "minimal" edits similar to our contrast sets, e.g., in image captioning (Shekhar et al., 2017), machine translation (Sennrich, 2017; Burlot and Yvon, 2017; Burlot et al., 2018), and language modeling (Marvin and Linzen, 2018; Warstadt et al., 2019). Minimal pairs of edits that perturb gender or racial attributes are also useful for evaluating social biases (Rudinger et al., 2018; Zhao et al., 2018; Lu et al., 2018). Our key contribution over this prior work is in grouping perturbed instances into a contrast set, for measuring local alignment of decision boundaries, along with our new, related resources. Additionally, rather than creating new data from scratch, contrast sets augment existing test examples to fill in systematic gaps. Thus contrast sets often require less effort to create, and they remain grounded in the original data distribution of some training set.

Since the initial publication of this paper, Shmidman et al. have further demonstrated the utility of contrast sets by applying these ideas to the evaluation of morphological disambiguation in Hebrew.

Recollecting Test Sets Recht et al. (2019) create new test sets for CIFAR and ImageNet by closely following the procedure used by the original datasets authors; Yadav and Bottou (2019) perform similar for MNIST. This line of work looks to evaluate whether *reusing* the exact same test set in numerous research papers causes the community to adaptively "overfit" its techniques to that test set. Our goal with contrast sets is different—we look to eliminate the biases in the *original annotation process* to better evaluate models. This cannot be accomplished by simply collecting more data because the new data will capture similar biases.

6 Conclusion

We presented a new annotation paradigm, based on long-standing ideas around contrastive examples, for constructing more rigorous test sets for NLP. Our procedure maintains most of the established processes for dataset creation but fills in some of the systematic gaps that are typically present in datasets. By shifting evaluations from accuracy on *i.i.d.* test sets to consistency on contrast sets, we can better examine whether models have learned the desired capabilities or simply captured the idiosyncrasies of a dataset. We created contrast sets for 10 NLP datasets and released this data as new evaluation benchmarks.

We recommend that future data collection efforts create contrast sets to provide more comprehensive evaluations for both existing and new NLP datasets. While we have created thousands of new test examples across a wide variety of datasets, we have only taken small steps towards the rigorous evaluations we would like to see in NLP. The last several years have given us dramatic modeling advancements; our evaluation methodologies and datasets need to see similar improvements.

References

Lars Ahrenberg. 2007. LinES: an English-Swedish parallel treebank. In *NODALIDA*.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task*

Papers, Belgium, Brussels. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *ACL*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *NAACL*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural Yes/No questions. In *NAACL*.

Michael Collins and James Brooks. 1995. Prepositional phrase attachment through a backed-off model. In *Third Workshop on Very Large Corpora*.

Pradeep Dasigi, Nelson F Liu, Ana Marasovic, Noah A Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *ACM AIES*.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *ICLR*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL*.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, Copenhagen, Denmark. Association for Computational Linguistics.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. In *ACL*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *EMNLP*.

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 *EMNLP*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *ACL*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL*.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. In *EMNLP*.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *EMNLP*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*.

Daniel Khashabi, Tushar Khot, and Ashish Sabhwawal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *EMNLP*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *EMNLP*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *KR*.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *EMNLP MRQA Workshop*.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018. Detecting and correcting for label shift with black box predictors. In *ICML*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *ACL*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn treebank. In *Computational Linguistics*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *EMNLP*.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *NAACL*.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *ACL*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *COLING*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An Improved Neural Baseline for Temporal Relation Extraction. In *EMNLP*.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A Multi-Axis Annotation Scheme for Event Temporal Relations. In *ACL*.

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

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

K.L. Pike. 1946. *Phonemics: A Technique for Reducing Languages to Writing*. Number v. 2 in Phonemics: A Technique for Reducing Languages to Writing. Summer Institute of Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In **SEM*.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *ICML*.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? Evaluating consistency of question-answering models. In *ACL*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018b. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*.

Manuela Sanguinetti and Cristina Bosco. 2015. PartTUT: The Turin university parallel treebank. In *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*.

Daniel Selsam, Matthew Lamm, Benedikt Bünz, Percy Liang, Leonardo de Moura, and David L. Dill. 2019. Learning a SAT solver from single-bit supervision. In *International Conference on Learning Representations*.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. In *EACL*.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *ACL*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! Find One mismatch between image and language caption. In *ACL*.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. In *Journal of Statistical Planning and Inference*.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Moshe Koppel, and Reut Tsarfaty. A novel challenge set for hebrew morphological disambiguation and diacritics restoration. In *Findings of EMNLP*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *LREC*.

Alane Suhr and Yoav Artzi. 2019. NLVR2 visual bias analysis. *arXiv preprint arXiv:1909.10411*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *ACL*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An open-domain dataset of qualitative relationship questions. 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 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In **SEM*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. In *EMNLP*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. In *TACL*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2019. BLiMP: A benchmark of linguistic minimal pairs for english. *arXiv preprint arXiv:1912.00582*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *ACL*.

Chhavi Yadav and Léon Bottou. 2019. Cold case: The lost MNIST digits. In *NeurIPS*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Amir Zeldes. 2017. The GUM corpus: Creating multi-layer resources in the classroom. In *LREC*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *ACL*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *EMNLP*.

A Dataset Details

Here, we provide details for the datasets that we build contrast sets for.

Natural Language Visual Reasoning 2 (NLVR2) Given a natural language sentence about two photographs, the task is to determine if the sentence is true (Suhr et al., 2019). The dataset has highly compositional language, e.g., *The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing*. To succeed at NLVR2, a model is supposed to be able to detect and count objects, recognize spatial relationships, and understand the natural language that describes these phenomena.

Internet Movie Database (IMDb) The task is to predict the sentiment (positive or negative) of a movie review (Maas et al., 2011). We use the same set of reviews from Kaushik et al. (2020) in order to analyze the differences between crowd-edited reviews and expert-edited reviews.

Temporal relation extraction (MATRES) The task is to determine what temporal relationship exists between two events, i.e., whether some event happened *before* or *after* another event (Ning et al., 2018). MATRES has events and temporal relations labeled for approximately 300 news articles. The event annotations are taken from the data provided in the TempEval3 workshop (UzZaman et al., 2013) and the temporal relations are re-annotated based on a multi-axis formalism. We assume that the events are given and only need to classify the relation label between them.

English UD Parsing We use a combination of four English treebanks (GUM, EWT, LinES, ParTUT) in the Universal Dependencies parsing framework, covering a range of genres. We focus on the problem of prepositional phrase attachment: whether the head of a prepositional phrase attaches to a verb or to some other dependent of the verb. We manually selected a small set of sentences from these treebanks that had potentially ambiguous attachments.

Reasoning about perspectives (PERSPECTRUM) Given a debate-worthy natural language claim, the task is to identify the set of relevant argumentative sentences that represent perspectives for/against the claim (Chen et al., 2019). We focus on the stance prediction sub-task: a binary prediction of whether a relevant perspective is for/against the given claim.

Discrete Reasoning Over Paragraphs (DROP) A reading comprehension dataset that requires numerical reasoning, e.g., adding, sorting, and counting numbers in paragraphs (Dua et al., 2019). In order to compute the consistency metric for the span answers of DROP, we report the average number of contrast sets in which F_1 for all instances is above 0.8.

QUOREF A reading comprehension task with span selection questions that require coreference resolution (Dasigi et al., 2019). In this dataset, most questions can be localized to a single event in the passage, and reference an argument in that event that is typically a pronoun or other anaphoric reference. Correctly answering the question requires resolving the pronoun. We use the same definition for consistency for QUOREFas we did for *DROP*.

Reasoning Over Paragraph Effects in Situations (ROPES) A reading comprehension dataset that requires applying knowledge from a background passage to new situations (Lin et al., 2019). This task has background paragraphs drawn mostly from science texts that describe causes and effects (e.g., that brightly colored flowers attract insects), and situations written by crowd workers that instantiate either the cause (e.g., bright colors) or the effect (e.g., attracting insects). Questions are written that query the application of the statements in the background paragraphs to the instantiated situation. Correctly answering the questions is intended to require understanding how free-form causal language can be understood and applied. We use the same consistency metric for ROPES as we did for DROP and QUOREF.

BoolQ A dataset of reading comprehension instances with Boolean (yes or no) answers (Clark et al., 2019). These questions were obtained from organic Google search queries and paired with paragraphs from Wikipedia pages that are labeled as sufficient to deduce the answer. As the questions are drawn from a distribution of what people search for on the internet, there is no clear set of "intended phenomena" in this data; it is an eclectic mix of different kinds of questions.

MC-TACO A dataset of reading comprehension questions about multiple temporal common-sense

phenomena (Zhou et al., 2019). Given a short paragraph (often a single sentence), a question, and a collection of candidate answers, the task is to determine which of the candidate answers are plausible. For example, the paragraph might describe a storm and the question might ask how long the storm lasted, with candidate answers ranging from seconds to weeks. This dataset is intended to test a system's knowledge of typical event durations, orderings, and frequency. As the paragraph does not contain the information necessary to answer the question, this dataset is largely a test of background (common sense) knowledge.

B Contrast Set Details

B.1 NLVR2

Text Perturbation Strategies We use the following text perturbation strategies for NLVR2:

- Perturbing quantifiers, e.g., *There is at least one dog → There is exactly one dog*.
- Perturbing numbers, e.g., *There is at least one dog → There are at least two dogs*.
- Perturbing entities, e.g., *There is at least one dog → There is at least one cat*.
- Perturbing properties of entities, e.g., *There is at least one yellow dog → There is at least one green dog*.

Image Perturbation Strategies For image perturbations, the annotators collected images that are perceptually and/or conceptually close to the hypothesized decision boundary, i.e., they represent a minimal change in some concrete aspect of the image. For example, for an image pair with 2 dogs on the left and 1 dog on the right and the sentence *There are more dogs on the left than the right*, a reasonable image change would be to replace the right-hand image with an image of two dogs.

Model We use LXMERT (Tan and Bansal, 2019) trained on the NLVR2 training dataset.

Contrast Set Statistics Five annotators created 983 perturbed instances that form 479 contrast sets. Annotation took approximately thirty seconds per textual perturbation and two minutes per image perturbation.

B.2 IMDb

Perturbation Strategies We minimally perturb reviews to flip the label while ensuring that the review remains coherent and factually consistent. Here, we provide example revisions:

Original (Negative): I had quite high hopes for this film, even though it got a bad review in the paper. I was extremely tolerant, and sat through the entire film. I felt quite sick by the end.
New (Positive): I had quite high hopes for this film, even though it got a bad review in the paper. I was extremely amused, and sat through the entire film. I felt quite happy by the end.
Original (Positive): This is the greatest film I saw in 2002, whereas I'm used to mainstream movies. It is rich and makes a beautiful artistic act from these 11 short films. From the technical info (the chosen directors), I feared it would have an anti-American basis, but ... it's a kind of (11 times) personal tribute. The weakest point comes from Y. Chahine : he does not manage to "swallow his pride" and considers this event as a well-merited punishment ... It is really the weakest part of the movie, but this testifies of a real freedom of speech for the whole piece.
New (Negative): This is the most horrendous film I saw in 2002, whereas I'm used to mainstream movies. It is low budgeted and makes a less than beautiful artistic act from these 11 short films. From the technical info (the chosen directors), I feared it would have an anti-American basis, but ... it's a kind of (11 times) the same. One of the weakest point comes from Y. Chahine : he does not manage to "swallow his pride" and considers this event as a well-merited punishment ... It is not the weakest part of the movie, but this testifies of a real freedom of speech for the whole piece.

Model We use the same BERT model setup and training data as Kaushik et al. (2020) which allows us to fairly compare the crowd and expert revisions.

Contrast Set Statistics We use 100 reviews from the validation set and 488 from the test set of Kaushik et al. (2020). Three annotators used approximately 70 hours to construct and validate the dataset.

B.3 MATRES

MATRES has three sections: TimeBank, AQUAINT, and Platinum, with the Platinum section serving as the test set. We use 239 instances (30% of the dataset) from Platinum.

Perturbation Strategies The annotators perturb one or more of the following aspects: appearance order in text, tense of verb(s), and temporal conjunction words. Below are example revisions:

- Colonel Collins **followed** a normal progression once she was **picked** as a NASA astronaut. (original sentence: "followed" is after "picked")
- Once Colonel Collins was **picked** as a NASA astronaut, she **followed** a normal progression. (appearance order change in text; "followed" is still after "picked")
- Colonel Collins **followed** a normal progression before she was **picked** as a NASA astronaut. (changed the temporal conjunction word from "once" to "before" and "followed" is now before "picked")

- Volleyball is a popular sport in the area, and more than 200 people were **watching** the game, the chief **said**. (original sentence: "watching" is before "said")
- Volleyball is a popular sport in the area, and more than 200 people would be **watching** the game, the chief **said**. (changed the verb tense: "watching" is after "said")

Model We use CogCompTime 2.0 (Ning et al., 2019).

Contrast Set Statistics Two annotators created 401 perturbed instances that form 239 contrast sets. The annotators used approximately 25 hours to construct and validate the dataset.

Analysis We recorded the perturbation strategy used for each example. 49% of the perturbations changed the "appearance order", 31% changed the "tense", 24% changed the "temporal conjunction words", and 10% had other changes. We double count the examples that have multiple perturbations. The model accuracy on the different perturbations is reported in the table below.

Perturbation Type	Accuracy
Overall	63.3%
Appearance Order	66.5%
Tense Change	61.8%
Temporal Conjunction	60.0%
Other Changes	61.8%

Table 4: Accuracy breakdown of the perturbation types for MATRES.

B.4 Syntactic Parsing

Perturbation Strategies The annotators perturbed noun phrases adjacent to prepositions (leaving the preposition unchanged). For example, *The clerics demanded talks with local US commanders → The clerics demanded talks with great urgency*. The different semantic content of the noun phrase changes the syntactic path from the preposition *with* to the parent word of the parent of the preposition; in the initial example, the parent is *commanders* and the grandparent is the noun *talks*; in the perturbed version, the grandparent is now the verb *demanded*.

Model We use a biaffine parser following the architecture of Dozat and Manning (2017) with ELMo embeddings (Peters et al., 2018), trained on the combination of the training sets for the treebanks that we drew test examples from (GUM, EWT, LinES, and ParTUT).

Contrast Set Statistics One annotator created 150 perturbed examples that form 150 contrast sets. 75 of the contrast sets consist of a sentence in which a prepositional phrase attaches to a verb, paired with an altered version where it attaches to a noun instead. The other 75 sentences were altered in the opposite direction.

Analysis The process of creating a perturbation for a syntactic parse is highly time-consuming. Only a small fraction of sentences in the test set could be altered in the desired way, even after filtering to find relevant syntactic structures and eliminate unambiguous prepositions (e.g. *of* always attaches to a noun modifying a noun, making it impossible to change the attachment without changing the preposition). Further, once a potentially ambiguous sentence was identified, annotators had to come up with an alternative noun phrase that sounded natural and did not require extensive changes to the structure of the sentence. They then had to re-annotate the relevant section of the sentence, which could include new POS tags, new UD word features, and new arc labels. On average, each perturbation took 10–15 minutes. Expanding the scope of this augmented dataset to cover other syntactic features, such as adjective scope, apposition versus conjunction, and other forms of clausal attachment, would allow for a significantly larger dataset but would require a large amount of annotator time. The very poor contrast consistency on our dataset (17.3%) suggests that this would be a worthwhile investment to create a more rigorous parsing evaluation.

Notably, the model's accuracy for predicting the target prepositions' grandparents in the original, unaltered tree (64.7%) is significantly lower than the model's accuracy for grandparents of all words (78.41%) and for grandparents of all prepositions (78.95%) in the original data. This indicates that these structures are already difficult for the parser due to structural ambiguity.

B.5 PERSPECTRUM

Perturbation Strategies The annotators perturbed examples in multiple steps. First, they created non-trivial negations of the claim, e.g., *Should we live in space? → Should we drop the ambition to live in space?*. Next, they labeled the perturbed claim with respect to each perspective. For example:

> **Claim:** Should we live in space?
> **Perspective:** Humanity in many ways defines itself through exploration and space is the next logical frontier.
> **Label:** True
>
> ---
>
> **Claim:** Should we drop the ambition to live in space?
> **Perspective:** Humanity in many ways defines itself through exploration and space is the next logical frontier.
> **Label:** False

Model   We use a ROBERTA model (Liu et al., 2019) finetuned on PERSPECTRUM following the training process from (Chen et al., 2019).

Contrast Set Statistics   The annotators created 217 perturbed instances that form 217 contrast sets. Each example took approximately three minutes to annotate: one minute for an annotator to negate each claim and one minute each for two separate annotators to adjudicate stance labels for each contrastive claim-perspective pair.

B.6   DROP

Perturbation Strategies   See Section 3 in the main text for details about our perturbation strategies.

Model   We use MTMSN (Hu et al., 2019), a DROP question answering model that is built on top of BERT Large (Devlin et al., 2019).

Contrast Set Statistics   The total size of the augmented test set is 947 examples and contains a total of 623 contrast sets. Three annotators used approximately 16 hours to construct and validate the dataset.

Analysis   We bucket 100 of the perturbed instances into the three categories of perturbations described in Section 3. For each subset, we evaluate MTMSN's performance and show the results in the Table below.

Perturbation Type	Frequency	Accuracy
Adding Compositional Steps	38%	67.5 F_1
Inversion of Semantics	37%	53.2 F_1
Re-ordering Events	25%	47.3 F_1

Table 5: Accuracy breakdown of the perturbation types for DROP.

B.7   QUOREF

Perturbation Strategies   We use the following perturbation strategies for QUOREF:

- Perturb questions whose answers are entities to instead make the answers a property of those entities, e.g., *Who hides their identity ... → What is the nationality of the person who hides their identity ....*
- Perturb questions to add compositionality, e.g., *What is the name of the person ... → What is the name of the father of the person ....*
- Add sentences between referring expressions and antecedents to the context paragraphs.
- Replace antecedents with less frequent named entities of the same type in the context paragraphs.

Model   We use XLNet-QA, the best model from Dasigi et al. (2019), which is a span extraction model built on top of XLNet (Yang et al., 2019).

Contrast Set Statistics   Four annotators created 700 instances that form 415 contrast sets. The mean contrast set size (including the original example) is $2.7(\pm 1.2)$. The annotators used approximately 35 hours to construct and validate the dataset.

B.8   ROPES

Perturbation Strategies   We use the following perturbation strategies for ROPES:

- Perturbing the background to have the opposite causes and effects or qualitative relation, e.g., *Gibberellins are hormones that cause the plant to grow → Gibberellins are hormones that cause the plant to stop growing.*
- Perturbing the situation to associate different entities with different instantiations of a certain cause or effect. For example, *Grey tree frogs live in wooded areas and are difficult to see when on tree trunks. Green tree frogs live in wetlands with lots of grass and tall plants. → Grey tree frogs live in wetlands areas and are difficult to see when on stormy days in the plants. Green tree frogs live in wetlands with lots of leaves to hide on.*
- Perturbing the situation to have more complex reasoning steps, e.g., *Sue put 2 cubes of sugar into her tea. Ann decided to use granulated sugar and added the same amount of sugar to her tea. → Sue has 2 cubes of sugar but Ann has the same amount of granulated sugar. They exchange the sugar to each other and put the sugar to their ice tea.*
- Perturbing the questions to have presuppositions that match the situation and background.

Model We use the best model from Lin et al. (2019), which is a span extraction model built on top of a RoBERTa model (Liu et al., 2019) that is first finetuned on RACE (Lai et al., 2017).

Contrast Set Statistics Two annotators created 974 perturbed instances which form 974 contrast sets. The annotators used approximately 65 hours to construct and validate the dataset.

B.9 BoolQ

Perturbation Strategies We use a diverse set of perturbations, including adjective, entity, and event changes. We show three representative examples below:

Paragraph: The Fate of the Furious premiered in Berlin on April 4, 2017, and was theatrically released in the United States on April 14, 2017, playing in 3D, IMAX 3D and 4DX internationally... A spinoff film starring Johnson and Statham's characters is scheduled for release in August 2019, while the ninth and tenth films are scheduled for releases on the years 2020 and 2021.
Question: Is "Fate and the Furious" the last movie?
Answer: False
New Question: Is "Fate and the Furious" the first of multiple movies?
New Answer: True
Perturbation Strategy: Adjective Change

Paragraph: Sanders played football primarily at cornerback, but also as a kick returner, punt returner, and occasionally wide receiver... An outfielder in baseball, he played professionally for the New York Yankees, the Atlanta Braves, the Cincinnati Reds and the San Francisco Giants, and participated in the 1992 World Series with the Braves.
Question: Did Deion Sanders ever win a world series?
Answer: False
New Question: Did Deion Sanders ever play in a world series?
New Answer: True
Perturbation strategy: Event Change

Paragraph: The White House is the official residence and workplace of the President of the United States. It is located at 1600 Pennsylvania Avenue NW in Washington, D.C. and has been the residence of every U.S. President since John Adams in 1800. The term is often used as a metonym for the president and his advisers.
Question: Does the president live in the White House?
Answer: True
New Question: Did George Washington live in the White House?
New Answer: False
Perturbation Strategy: Entity Change

Model We use ROBERTA base and follow the standard finetuning process from Liu et al. (2019).

Contrast Set Statistics The annotators created 339 perturbed questions generated that form 70 contrast sets. One annotator created the dataset and a separate annotator verified it. This entire process took approximately 16 hours.

B.10 MC-TACO

Perturbation Strategies The main goal when perturbing MC-TACO questions is to retain a similar question that requires the same temporal knowledge to answer, while there are additional constraints with slightly different related context that changes the answers. We also modified the answers accordingly to make sure the question has a combination of plausible and implausible candidates.

Model We use the best baseline model from the original paper (Zhou et al., 2019) which is based on ROBERTA$_{base}$ (Liu et al., 2019).

Contrast Set Statistics The annotators created 646 perturbed question-answer pairs that form 646 contrast sets. Two annotators used approximately 12 hours to construct and validate the dataset.