A Dataset for Tracking Entities in Open Domain Procedural Text

Niket Tandon[1] **Keisuke Sakaguchi**[1] **Bhavana Dalvi Mishra**[1] **Dheeraj Rajagopal** [2]
Peter Clark[1] **Michal Guerquin**[1] **Kyle Richardson**[1] **Eduard Hovy**[2]
[1]Allen Institute for Artificial Intelligence, Seattle, WA
{nikett, keisukes, bhavanad, peterc, michalg, kyler}@allenai.org
[2]Department of Computer Science, CMU
{dheeraj, hovy}@cs.cmu.edu

Abstract

We present the first dataset for tracking state changes in procedural text from arbitrary domains by using an *unrestricted* (open) vocabulary. For example, in a text describing fog removal using potatoes, a `car window` may transition between being *foggy, sticky, opaque*, and *clear*. Previous formulations of this task provide the text and entities involved, and ask how those entities change for just a small, pre-defined set of attributes (e.g., location), limiting their fidelity. Our solution is a new task formulation where given just a procedural text as input, the task is to generate a set of state change tuples *(entity, attribute, before-state, after-state)* for each step, where the entity, attribute, and state values must be predicted from an open vocabulary. Using crowdsourcing, we create OPENPI[1], a high-quality (91.5% coverage as judged by humans and completely vetted), and large-scale dataset comprising 29,928 state changes over 4,050 sentences from 810 procedural real-world paragraphs from WikiHow.com. A current state-of-the-art generation model on this task achieves 16.1% F1 based on BLEU metric, leaving enough room for novel model architectures.

1 Introduction

By one estimate, only about 12% of what we understand from text is expressed explicitly (Graesser, 1981). This is especially apparent in text about actions where the effects of actions are left unstated. Humans fill that gap easily with their commonsense but machines need to model these effects in the form of state changes. For example, when *a potato is rubbed on a car window* (to defog it), then the unstated effects of this action are the following state changes: *windows becomes sticky, opaque, and the potato becomes dirty*, etc. These changes can be tracked across the paragraph. An exemplary use case of text with actions is procedural text (recipes, how-to guides, etc.) where modeling such state changes helps in various reasoning-based end tasks, e.g. automatic execution of biology experiments (Mysore et al., 2019), cooking recipes (Bollini et al., 2012) and everyday activities (Yang and Nyberg, 2015).

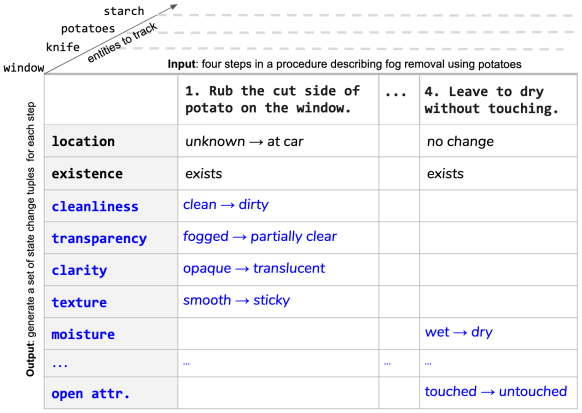

Figure 1: Previous formulations of the state tracking task are restricted to a small, fixed set of pre-defined state change types that limits its fidelity to model real-world procedures (they cannot cover the blue part in this procedure comprising four steps). Our solution is a new task formulation to track an unrestricted (open) set of state changes (additionally covering blue).

While there has been great progress in tracking entity states in scientific processes (Dalvi et al., 2018), tracking ingredients in cooking recipes (Bosselut et al., 2018), and tracking the emotional reactions and motivations of characters in simple stories (Rashkin et al., 2018), prior tasks are restricted to a fixed, small set of state change types thus covering only a small fraction of the entire world state. Figure 1 illustrates this for a real-world procedure "How to Keep Car Windows Fog Free Using a Potato". Existing datasets such as ProPara (Dalvi et al., 2018) only model the *existence* and *location* attributes, limiting the fidelity with which they model the world. Specifically:

- Attributes from domain-specific datasets such

[1]Download OPENPI at https://allenai.org/data/openpi

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6408–6417,
November 16–20, 2020.

as ProPara (Dalvi et al., 2018) and Recipes (Bosselut et al., 2018) together, only cover ∼40% of the state changes that people typically mention when describing state changes in real-world paragraphs from WikiHow (§2.1).

- The set of attributes that people naturally use to describe state changes is large, and hence hard to pre-enumerate ahead of time (especially when the target domain is unknown). Even a comprehensive list of popular attributes failed to cover 20% of those used in practice (§4.2).
- The dominant approach in existing datasets is to assume that changing entities are mentioned as spans in the procedural text. However, in unconstrained human descriptions of changes, ∼40% of the referred-to entities were unmentioned in the text (e.g., the knife and cutting board in several cooking recipes) (§4.4).

Addressing these limitations, our solution is a new task formulation to track an unrestricted (open) set of state changes: Rather than provide the text and entities, and ask how those entities change for a pre-defined set of attributes at each step, we instead provide just the input text, and ask for the set of state changes at each step, each describing the before and after values of an attribute of an entity in the form ($attribute$ *of* $entity$ *was* $value_{\text{before}}$ *before and* $value_{\text{after}}$ *afterwards.*). Importantly, the vocabularies for attributes, entities, and values is open (not pre-defined). Our contributions are:

(i) we introduce a novel task of tracking an unrestricted (open) set of state change types (§2).

(ii) we create a large-scale (∼30K state changes), high-quality ∼ 91.5% coverage and human vetted) crowdsourced annotated dataset OPENPI, from a general domain text serving as training dataset for this task (§4).

(iii) we establish a strong generation baseline demonstrating the difficulty of this task (§5), and present an error analysis suggesting avenues for future research (§6.3).

2 Proposed Task: OPENPI

From a procedural paragraph with sentences (i.e., steps) $step_1 \ldots step_K$, construct K data points, one per step.

Input: As input we are given a procedural text comprising current step $step_i$ as query and all past step as context $step_1 \cdots step_{i-1}$. We denote the input as $x = (x_q, x_c)$, where x_q is the step for which we need the state changes (i.e. the query) and x_c is the context.

Here, we use the common assumption (Dalvi et al., 2018) that the steps in procedural text are ordered such that the context required for $step_i$ is mentioned in $step_1 \cdots step_{i-1}$.

Output: The output is a set of zero or more state changes $\boldsymbol{y}$= $\{y_i\}$. A state change y_i is of the form: $attr$ *of* ent *was* val_{pre} *before and* val_{post} *afterwards*

Here, $attr$ is the attribute or state change type, and ent is the changed entity. val_{pre} is the precondition (i.e., the state value before), and val_{post} is the postcondition (i.e., the state value afterwards). Pre/ postcondition $adj\_or\_relp(y_i^{pre})$ can be an adjectival phrase or a relational phrase. In this task, $attr$, ent, val_{pre} and val_{post} are open form text i.e. they are not tied to any fixed, constrained vocabulary.

Example: Consider the running example: x=(context: *The window of your car is foggy*, query: *Rub half potato on the window*). Then, $\{y\}$ = { *transparency of window was fogged before and partially clear afterwards, stickiness of window was smooth before and sticky afterwards* }. In y_1, $attr$ = *transparency*, ent = `window`, val_{pre} = *fogged* and val_{post} = *partially clear*

2.1 Unique Challenges

OPENPI has two unique challenges that are not found in any existing state change dataset.

- **Variable size, low-specificity output**: (Jas and Parikh, 2015) introduce the notion of image specificity which measures the amount of variance in multiple viable descriptions of the same image (typically, each image has exactly K descriptions from K annotators). Low specificity implies very different descriptions that are not mere re-phrasings. In OPENPI the output $\boldsymbol{y}$ has low-specificity (low specificity is also called high complexity output). To achieve low specificity outputs, existing methods learn to generate diverse responses by sampling different keywords and using a reinforcement learning approach for training (Gao et al., 2019) or use a diverse beam search (Vijayakumar et al., 2018) based approach on a typical encoder to decode diverse outputs. However, they all assume that the output set size is fixed to K (typically each

Input x	**Output** $\boldsymbol{y}$
Apply insecticide to peonies.	*the location of insecticide was in bottle before and on peonies afterwards.* *the health of bugs were healthy before and dying afterwards.*
Dip the peony flowers in water.	*the moisture of flowers was dry before and wet afterwards.* *the cleanliness of peonies were dirty before and clean afterwards.*
Stop ants from climb.. use trap	*the organization of trap was disassembled before and assembled after..* *the well being of plants were troubled before and healthy afterwards.*
Combine apricots, .. in blender.	*the location of apricots was on counter before and in blender afterwards.* *the state of ingredients were separate before and combined afterwards.* *the weight of blender was light before and heavy afterwards.*
Add oil until dressing thick.	*the state of ingredients were separate before and combined afterwards.* *the location of oil was on counter before and in blender afterwards.*
Stir in the basil.	*the location of dressing was in blender before and on serving plate..* *the location of basil was outside blender before and in blender afterwards.* *the weight of blender was heavy before and light afterwards.*

Table 1: Examples of the task based on our dataset. The input x comprises a ***query*** x_q and a context x_c (past sentences before this step in the paragraph– not shown due to limited space). The output is a set $\boldsymbol{y}$ of pre and postconditions. The paragraphs in this table are: above (how to clean oven) and below (cooking recipe).

sample is annotated by exactly K annotators). In our case, however, the number of items in $\boldsymbol{y}$ is variable, making these existing solutions inapplicable.

- **Open vocabulary**: In OPENPI $attr$, ent, val_{pre} and val_{post} are not restricted to any fixed, small vocabulary. Previous task formulations such as (Bosselut et al., 2018; Dalvi et al., 2018), made the assumption that ent is given, $attr$ is from a vocabulary of less than 10 classes, and val_{pre} or val_{post} are either from a small external vocabulary or a span in x[2]. In contrast, in OPENPI, the entities may not be present in the sentence or even the context, and the state change types and values can come from a rather open vocabulary. This openness brings a variety of challenges: (i) presupposed entities: these are entities that are not present in x and perceived through background knowledge, (ii) zero shot learning: during inference on a previously unseen domain, there are previously unseen attributes, entities, and state change types. This makes the problem very challenging and places this task in a novel setting (see §3.1)

[2]We matched an exhaustive list of synonyms of existing attributes from existing datasets ProPara and Recipes: *existence, location, temperature, composition, cleanliness, cookedness, shape, rotation, accessibility* and found that only ∼40% of the attributes in OPENPI are covered by these (however, these datasets cannot cover the open vocabulary of entities and attribute values)

Task	Vocab.	Specificity	Output size
Story CSK	open	high	fixed
ProPara	closed	high	fixed
Recipes Task	closed	low	fixed
ALFRED	closed	high	fixed
VirtualHome	closed	high	fixed
OpenPI	open	low	variable

Table 2: Comparison of our dataset to existing datasets

3 Related Work

Tracking state changes: Procedural text understanding addresses the task of tracking entity states throughout the text (Bosselut et al., 2018; Henaff et al., 2017). This ability is an important part of

text understanding. While syntactic parsing methods such as AMR (abstract meaning representation) (Banarescu et al., 2013) represent "who did what to whom" by uncovering stated facts, tracking entity states uncovers unstated facts such as how ingredients change during a recipe.

Datasets with closed state changes: The bAbI dataset (Weston et al., 2015) includes questions about objects moved throughout a paragraph, using machine-generated language over a deterministic domain with a small lexicon. The SCoNE dataset (Long et al., 2016) contains paragraphs describing a changing world state in three synthetic, deterministic domains. However, approaches developed using synthetic data often fail to handle the inherent complexity in language when applied to organic, real-world data (Hermann et al., 2015; Winograd, 1972). The ProPara dataset (Dalvi et al., 2018) contains three state changes *(create, destroy, move)* for natural text describing scientific procedures. Other domain specific datasets include recipe domain (Bosselut et al., 2018), and biology experiments (Mysore et al., 2019). These datasets contain a small, closed set of state change types that are relevant to a specific domain. Our dataset is general domain, and to accommodate this generality we have an open vocabulary of state changes.

Datasets with open state changes: (Isola et al., 2015) propose manually defined antonymous adjective pairs (big, small) to define transformations in images, and this was an inspiration for us to use adjectives as open state changes in OPENPI Knowledge bases such as ConceptNet (Speer and Havasi, 2013) and ATOMIC (Sap et al., 2019) contain (open) pre-conditions and post-conditions but they are agnostic to context. Context plays a role when dealing with a large number of state changes types e.g., if "a stone hits a glass" then the glass would break but this is not the case if "a soft toy or a sound wave hits the glass". Our dataset contains context information, an important training signal for neural models.

Current knowledge bases (such as ATOMIC) contain social rather than physical effects. As a result, generation models trained on these knowledge bases incorrectly force the effects to be social. For example, COMET (Bosselut et al., 2019), trained on ATOMIC data, when applied on "Cans are tied together and transported to a recycling center", incorrectly predicts[3] *person goes to recycle center, Person needs to be arrested → Person is arrested, gets dirty*.

3.1 Positioning OPENPI

Figure 2.1 projects existing tasks and models along two different dimensions (open vocabulary, and variable-size low-specificity). We find that models bottom-left quadrant represents majority of the existing work on state changes such as ProPara (Dalvi et al., 2018) and bAbI (Weston et al., 2016)) in NLP community, and ALFRED (Shridhar et al., 2019) and VirtualHome (Puig et al., 2018) in Computer Vision. Correspondingly many models exist in that space ((Tandon et al., 2018), (Bosselut et al., 2018), (Henaff et al., 2017)). Very few models exist that can predict either open vocab (Rashkin et al., 2018), or variable size output (Bosselut et al., 2018). However, no existing task has both open vocabulary and variable-size low specificity– placing OPENPI in a novel space.

4 Dataset

4.1 Data Collection

We set up a crowdsourcing task on Amazon Mechanical Turk where the annotators author the $\boldsymbol{y}= \{y_i\}$ for every sentence of a wikihow.com article, filling in a sentence template for each y_i as a guide. WikiHow contains a wide variety of goals (e.g., *how to wash dishes*) broken down into steps with detailed descriptions and pictorial illustrations, spanning across 19 categories. We selected a diverse subset of six popular categories and focus on action-oriented articles[4].

For a given WikiHow article, annotators were asked to describe up to six state changes for each step ($0 \leq |\boldsymbol{y}| \leq 6$), and were paid fairly[5]. Each state change description consists of precondition (y_i^{pre}), postcondition (y_i^{post}), and the (physical) attribute. Restricting the annotators to a template for state change described in §2, yields much better quality than free-form. This was a pragmatic

[3]Manually inspecting the 45 predictions made by COMET on this sentence, we found only one partially correct prediction that the human has to get to the recycle center before.

[4]We exclude WikiHow articles with steps containing stative verbs such as *know, see, want*, etc., and remove articles with too few (less than 4) or too many steps (7 or more). The selected categories are in Table 3.

[5]We set the reward to be $0.07 for each of the first three state changes, and $0.14 for each of the additional three state changes in order to encourage workers to write as many state changes as possible. All annotators met the following prerequisites as a minimum qualification: (1) 5K previous HITs approvals, (2) 99% or higher approval rate, (3) location is US, UK, CA, AU, or NZ.

choice, to encourage Turkers to give a complete description but not add extra noise. In an earlier pilot, we tried upto 10 changes but Turkers found the task too difficult and complained. Six empirically resulted in the best level of agreement and completeness among annotations, while also retaining diversity.

The annotators were encouraged (but not required) to pick from a pre-defined vocabulary of 51 WordNet derived attributes.

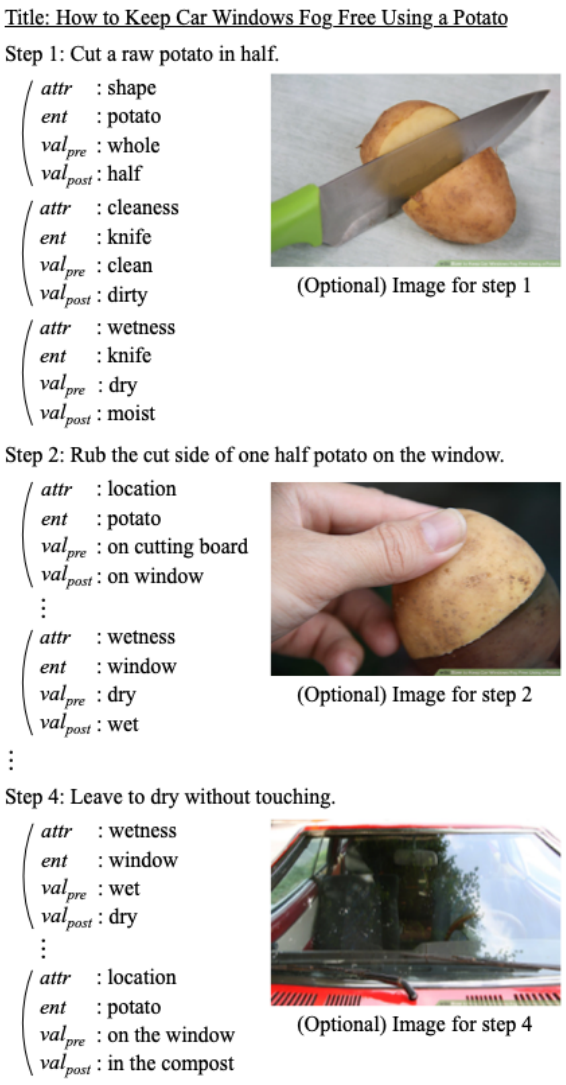

Figure 2: Data collection procedure: Crowdworkers are shown the article title, step descriptions and optionally the corresponding image, and asked to write up to six state changes (y_i^{pre}, y_i^{post}, *attr*) per step. See the appendix for a sample of the annotation task.

We performed two sets of annotations for every article, one where the annotators see the pictorial illustration of a step and one without. Visuals helped the annotators to provide more state changes (e.g., the color of cut potato turns gray). In total, one article is annotated four times (two turkers each for with and without images)– making the cost of annotation $3.6 in average per article. See Figure 2 for an example of the annotation procedure.

After collecting the data, we cleaned up the state changes by asking three crowd workers if each state change is valid or not with the same annotation setting as data collection (e.g., with or without visual illustration). We discarded state changes that did not get the agreement by the majority (2 or more workers). With this cleaning step, the total number of state changes changed from 33,065 to 29,928.

The small number of errors encountered during vetting fell into five categories:

- (∼45% of the errors) Obscure attributes/ values, e.g., *state of clubhouse was spoken of before.*
- (∼20%) State change of future steps, e.g., *Prepare the pot* → *location of veggies in pot*
- (∼15%) Mismatch of attribute and value: *shape of lemon was solid*
- (∼10%) State change of the reader, not the actor: *knowledge of you becomes aware*
- (∼10%) Factual errors: annotated change does not occur or tautologously refers to the action.

4.2 Dataset statistics

The resulting OPENPI dataset comprises 29,928 state changes over 4,050 sentences from 810 WikiHow articles. Of these, 15,445 (4.3 per step) state changes were obtained from the *with images* setting and 14,483 (3.8 per step) from *without images*, indicating that the additional visual modality helped workers to come up with more state changes (e.g., the color of cut potato turns gray). These WikiHow articles were from six categories, see Table 3. The number of state changes in a category depends on the density of entities and their changes e.g., cooking related articles include multiple ingredients and instruments that undergo state changes.

Two thirds of the state changes are adjective phrases (avg. length 1.07 words) and the remaining one third are relational phrases (avg. length 2.36 words). Attributes, entities, adjective phrases, relational phrases all follow a power-law like distribution. The most popular adjectives were {*dry, empty, clean, wet, dirty, full, heavier, lighter, hot, whole, cool, cold*}, and the most popular relational phrases were location-indicating prepositions. About 20% of the attributes are present in 80% of the data.

WikiHow cat.	# para	$\|y\|$	w/ img	w/o
Food & Entertain	197	9942	5399	4543
Home & Garden	199	6961	3758	3203
Hobbies & Craft	193	4766	2375	2391
Sports & Fitness	95	3361	1662	1699
Cars & Vehicles	43	1656	818	838
Health	77	3036	1433	1603
All	858	29928	15445	14483

Table 3: Basic statistics of the OPENPI dataset: the articles' WikiHow category, the number of WikiHow articles (i.e., paragraphs) in each category and number of state changes $|y|$ (total), and data collected using with, and without image setting).

The long tail of the remaining 80% attributes indicates why open attributes are important. As similar attributes can be expressed differently in text e.g., *wetness* and *moisture*, we analyzed a few data points to observe a large agreement between annotators in choosing attributes (the average size of attribute clusters was only 1.2).

We split the data into training, development, and test sets. To evaluate model generalization, all the annotated articles in the Health category are marked as out-of-domain and placed (only) in the test set. All the remaining annotated articles are randomly assigned to one of the splits. The resulting training set has 23,869 state changes (3,216 instances because one instance comprises $|y|$ state changes), dev set has 1,811 (274 instances), and test set has 4,248 (160 instances in domain, and 394 instances out-of-domain "Health").

4.3 Dataset quality

We measure the quality (coverage) of the dataset by asking a human judge whether there is any new state change they can add. The judge added only 8.5% new state changes. This suggests that OPENPI has a high coverage of ∼91.5%, and a very high precision because of vetting.

These additions fell into four categories:

- (∼40% of additions) Indirect effect was missed, e.g., *Place in freezer* → (existing) *food cooler*, (added) *food container cooler*
- (∼35%) Extra dimension of change (attribute) missed, e.g., (added) Change in texture, organization, open/closed state.
- (∼20%) Addition is a rewording hence not helpful e.g., *cleanliness of windshield*, (added) *clarity of windshield*
- (∼5%) Addition is incorrect/obscure.

4.4 Quantifying the reasoning challenges

Presupposed entities: About 61% of the entities in our development set are mentioned as spans in the context and paragraph, while the remaining 40% are unmentioned entities. About 35% of the unmentioned entities were derivatives of mentioned entities, i.e. synonym, hypernym-hyponym, or part-whole. The remaining 65% were presupposed (assumed) entities, e.g., containers of mentioned entities, surfaces, cooking instruments.

Open attributes: 78.9% of the examples contain the 51 predefined attributes that the annotators were supplied. The remaining examples contain 577 Turk authored open attributes and many of these are difficult to anticipate, e.g., cookedness, tightness, saltiness. This makes up a long tail distribution of an open vocabulary of attributes.

Zero-shot learning: The test-set contains: 1) paragraphs from five categories covered in the training set, 2) paragraphs from `Health` category for which there is no training data, to test zero-shot learning. `Health` test-subset is particularly challenging with 55% unmentioned entities (40% otherwise) and 33% unseen attributes (18% otherwise).

Variable size, low specificity output: A system needs to decide relevant entities and attributes would be relevant and generate possibly varying number of state changes for different steps. The dev set has on average seven state changes per step, and 3% of the steps have no state change.

5 Model

OPENPI dataset poses unique challenges including presupposed entities, open attributes, zero-shot learning and variable-size, low specificity output (see Section 4.4). These challenges make it difficult to apply existing entity tracking methods like ProStruct (Tandon et al., 2018), EntNet (Henaff et al., 2017), NPN (Bosselut et al., 2018) without making significant changes to either the model or the task. E.g., the commonsense constraints in ProStruct do not scale with a large number of attributes, and EntNet is not suitable for a set output.

OPENPI is well-suited for a generation model because the output *attr of ent was* val_{pre} *before and* val_{post} *afterwards* must be predicted using an open vocabulary. Therefore, as our baseline, we use the state-of-the-art pre-trained language model, GPT-2 (Radford et al., 2019), and fine-tune it for

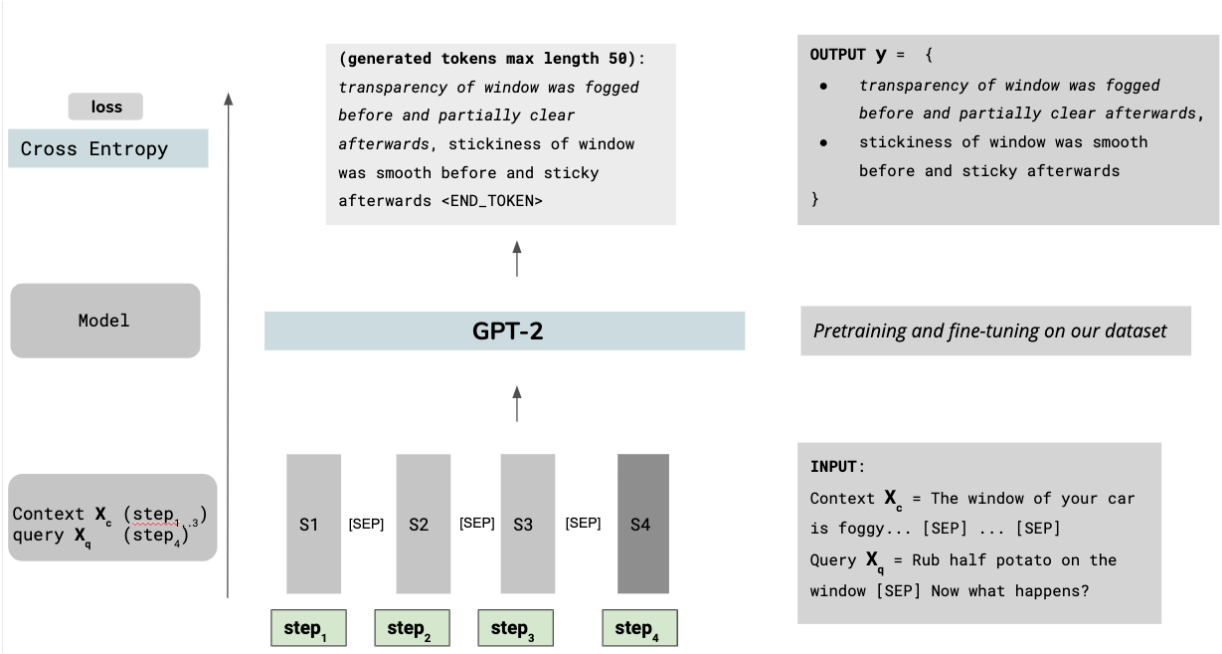

Figure 3: Our GPT-2 based model for OPENPI

OPENPI task. The model takes as input a special [SEP] token separated x_c and x_q. The output is expected to be a set $\boldsymbol{y}$ of variable size. As noted in §2.1, existing methods do not produce a variable size, low-specificity output. Instead we train the model to generate a long sequence of comma separated y_i. If there are no changes i.e., $|\boldsymbol{y}| = 0$, then we set $\boldsymbol{y} = \{\textit{there will be no change}\}$.

Figure 3 shows the model architecture. During decoding, we sample y_i as a sequence of output tokens generated by the model. The generation output accounts for all aspects of the state change - the attribute, entity, and before, after values.

6 Experiments

6.1 Metrics

To measure the performance on OPENPI we compare the predicted set $\boldsymbol{y}$ and gold set $\boldsymbol{y}$*, for every point x. Precision for a data point x is computed based on the best matching gold state change for each predicted state change i.e., $P(x) = \frac{1}{2}\sum_{y \in \boldsymbol{y}} max_{y*}\, O(y*^{pre}, y^{pre}) + O(y*^{post}, y^{post})$. Similarly, recall is based on the best matching predicted state change for each gold state change i.e. $R(x) = \frac{1}{2}\sum_{y* \in \boldsymbol{y}^*} max_{y}\, O(y*^{pre}, y^{pre}) + O(y*^{post}, y^{post})$. The string overlap function $O(.)$ can use any of the standard generation metrics: exact match, BLEU, METEOR or ROUGE[6]. We report micro-averaged precision, recall, F1 scores for different choices of $O(.)$.

We remove template words before string comparison to avoid inflating scores for template words. We did not perform facet-based evaluation of the templated output for two reasons. Firstly, while it might seem when computing overlaps of gold and predicted state changes as two long strings, BLEU or ROUGE may accidentally see an overlap when there was none. That is unlikely in practice because the entities, attributes, and values are quite distinct and scoring accidental overlaps is uncommon. Secondly, our evaluation metric (F1, precision, recall) matches a list of predictions against a list of gold references. It is unclear how to compute F1 over individual facets that requires the best match based on all facets as tuple.

We also found that when manually evaluating on ∼200 dev datapoints, the score was systematically a few (∼10%) points higher than BLEU, while the trends and model rankings remained the same, indicating robustness of the automatic metric.

Therefore, the proposed metric aligns with human evaluation, and is able to use existing generation metrics thereby simplifying evaluation, allowing easier reproducibility.

6.2 Evaluation

We evaluate state of the art generation model GPT-2 on OPENPI dataset. As mentioned in Section 4.2,

[6] github.com/allenai/abductive-commonsense-reasoning

OPENPI consists of two kinds of annotations: with-images (turkers were shown images along with text for each step of the procedure) and without-image (turkers looked at only text to predict state changes). GPT-2 gets to see only text as input but the state changes it has to predict are different depending on the setting. Table 4 reports P, R and F1 when GPT-2 model is tested on different subsets.

The GPT-2 model struggles to predict the right set of state changes indicating that the task is hard. Challenges include lexical variation on entities (in context vs. in gold), unseen categories, limited context for the initial sentences in the paragraph an so on. Detailed error analysis is presented in §6.3.

	F1 based on		
	Exact	BLEU	ROUGE
with-image	5.1	14.3	29.1
without-image	3.6	13.4	28.2
Entire dataset	4.3	16.1	32.4

Table 4: GPT-2 on OpenPI, and its sub-categories.

Models	BLEU scores		
	P	R	F1
seen category	25.1	18.4	17.1
unseen categories	24.4	17.4	15.7

Table 5: GPT-2 on topics seen, unseen during training.

OPENPI testset comprises of both unseen and seen categories, and we report BLEU separately on these subsets. Results from table 5 presents an encouraging result that GPT-2 generalizes to unseen topics even though the scores on seen categories is understandably a little higher (F1 of 17.1 for seen category vs 15.7 for unseen categories).

6.3 Error analysis

To better understand model shortcomings, the error types in dev predictions are illustrated (Table 6).

1. Wrong attribute ($attr(y_i)$): In 51% state changes produced by the GPT-2 model, predicted attribute is incorrect. Often (∼20% of cases) predicted attribute is *state*, i.e. the model couldn't name the attribute.

Gold:	wetness of potatoes was wet before, and dry after
Pred:	state of potatoes was wet before, and dry after

Error type	freq	%
Wrong attribute	826	51
Wrong entity	964	59
Wrong adjective	989	41
Wrong relation phrase	456	17
Any of the above	1,622	100

Table 6: Error types in 1,811 dev predictions. One state change prediction can have multiple error types.

2. Wrong entity ($ent(y_i^{pre})$): The model predicted incorrect entity 59% of the times. For 32% of the entity errors, the gold entity was unmentioned in the input text.

   (i) Entities present as span (68%): Typically, a related but not same entity is predicted:

G:	..furniture was worn out before, and renewed after
P:	..chairs was dirty before, and clean after

   (ii) Derivable entities: (3%) These entities are typically a lexical variation of the entities in the paragraph. E.g., *spray paint silk floral arrangement to change color or freshen its hue*, the model predicted

G:	..plant was dry before, and wet after
P:	..cloth was dry before, and wet after

   The following example also mentions a derivable entity and both gold and prediction are imply the same but it is difficult to automatically check that. E.g., *Keep the craft steady as others board.*

G:	stability of boat was rocking ... steadied after
P:	stability of craft was wobbling ... steady after

   (iii) Unmentioned entities: (29%). These types of errors are very difficult to overcome because the entities are typically not mentioned at all in the generated output. For instance in the following, loser and rider both refer to the same person in the text,

G:	..loser was alive before, and dead after
P:	..rider was alive before, and killed after

   In about 20% of such erroneous predictions, the model predicted the $adj(y_i^{pre})$ correctly. This may be because attribute is a good indicator of the adjectives.

3. Wrong $adj(y_i^{pre})$: (41%) The model pre-

dicts incorrect adjectives, such that in some cases the erroneous adjectives might not apply to the given entity, or the adjective values are swapped between pre and post condition. An example is shown below:

G	..curtains was white before, and painted after
P:	..double curtains was colorless ... colorful after

4. Wrong $relp(y_i^{pre})$ (17%): We find that relational phrases are very hard for the model currently. 184 out of 210 relational state changes predicted by the model have incorrect relational phrase. We believe that this poses a challenging research problem for future models.

G:	knowledge of animals was absent ... present after
P:	details afterwards was ignored ... discussed after

5. Length of the context plays an important role. Without any context (e.g., for the first step), the model gets a low accuracy of 8.3%.

7 Conclusion

We presented the first dataset to track entities in open domain procedural text. To this end, we crowdsourced a large, high-quality dataset with examples for this task. We also established a strong generation baseline highlighting the difficulty of this task. As future work, we will explore more sophisticated models that can address the highlighted shortcomings of the current model. An exciting direction is to leverage visuals of each step to deal with unmentioned entities and indirect effects.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *LAW@ACL*.

Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. 2012. Interpreting and executing recipes with a cooking robot. In *ISER*.

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2018. Simulating action dynamics with neural process networks. *ICLR*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: A challenge dataset and models for process comprehension. *NAACL*.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019. Generating multiple diverse responses for short-text conversation. In *AAAI*.

Arthur C. Graesser. 1981. Prose comprehension beyond the word. In *Springer*.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *ICLR*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. 2015. Discovering states and transformations in image collections. *CVPR*.

Mainak Jas and Devi Parikh. 2015. Image specificity. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2727–2736.

Reginald Long, Panupong Pasupat, and Percy Liang. 2016. Simpler context-dependent logical forms via model projections. In *ACL*.

Sheshera Mysore, Zach Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. *arXiv preprint arXiv:1905.06939*.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8494–8502.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *ACL*.

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

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2019. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. *ArXiv*, abs/1912.01734.

Robyn Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*.

Niket Tandon, Bhavana Dalvi Mishra, Joel Grus, Wentau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. *EMNLP*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. *AAAI*.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. *ICLR*, abs/1502.05698.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Zi Yang and Eric Nyberg. 2015. Leveraging procedural knowledge for task-oriented search. In *SIGIR*.