ABDUCTIVE COMMONSENSE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Abductive reasoning is inference to the most plausible explanation. For example, if Jenny finds her house in a mess when she returns from work, and remembers that she left a window open, she can hypothesize that a thief broke into her house and caused the mess, as the most plausible explanation. While abduction has long been considered to be at the core of how people interpret and read between the lines in natural language (Hobbs et al., 1988), there has been relatively little research in support of abductive natural language inference and generation.

We present the first study that investigates the viability of language-based abductive reasoning. We introduce a challenge dataset, ART, that consists of over 20k commonsense narrative contexts and 200k explanations. Based on this dataset, we conceptualize two new tasks – (i) Abductive NLI: a multiple-choice question answering task for choosing the more likely explanation, and (ii) Abductive NLG: a conditional generation task for explaining given observations in natural language.

On Abductive NLI, the best model achieves 68.9% accuracy, well below human performance of 91.4%. On Abductive NLG, the current best language generators struggle even more, as they lack reasoning capabilities that are trivial for humans. Our analysis leads to new insights into the types of reasoning that deep pre-trained language models fail to perform—despite their strong performance on the related but more narrowly defined task of entailment NLI—pointing to interesting avenues for future research.

1 INTRODUCTION

The brain is an abduction machine, continuously trying to prove abductively that the observables in its environment constitute a coherent situation.

– Jerry Hobbs, ACL 2013 Lifetime Achievement Award

Abductive reasoning is inference to the most plausible explanation for incomplete observations Peirce (1965a). Figure 1 illustrates an example. Given the incomplete observations about the world that $O_1$: “Jenny cleaned her house and went to work, leaving the window just a crack open.” and sometime later $O_2$: “When Jenny returned home, she saw her house was a mess.”, we can hypothesize different potential explanations and reason about which is the most likely. We can readily rule out $H_3$ since it fails to justify the observation $O_2$. While $H_1$ and $H_2$ are both plausible, the most likely explanation based on commonsense is $H_1$ as $H_2$ is somewhat implausible given $O_1$.

One crucial point Peirce makes about abductive reasoning is that abduction is “the only logical operation which introduces any new ideas”, which contrasts with other types of inference such as entailment, that focuses on inferring only such information that is already provided in the premise. Abductive reasoning has long been considered to be at the core of understanding narratives (Hobbs et al., 1988), reading between the lines (Norvig, 1987; Charniak & Shimony, 1990), reasoning about everyday situations (Peirce, 1965b; Andersen, 1973), and counterfactual reasoning (Pearl, 2002; Pearl & Mackenzie, 2018). Despite the broad recognition of its importance, however, the study of abductive reasoning in narrative text has very rarely appeared in the NLP literature, in large part because most previous work on abductive reasoning has focused on formal logic, which has proven to be too rigid to generalize to the full complexity of natural language.

1The full transcript of his award speech is available at https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00171
A thief broke into the house by pulling open the window. The bird got stuck inside the house, flew around while trying to escape, and made a mess. When Jenny returned home she saw that her house was a mess. Formally, the task requires to maximize $P(h^+ | O_1, O_2)$. Our analysis leads to insights into the types of reasoning that deep pre-trained language models fail to perform — despite their strong performance on the closely related but different task of entailment NLI — pointing to future research directions.

2 TASK DEFINITION

**Abductive Natural Language Inference** We formulate $\alpha$NLI as multiple choice problems consisting of a pair of observations as context and a pair of hypothesis choices. Each instance in ART is defined as follows:

- $O_1$: The observation at time $t_1$.
- $O_2$: The observation at time $t_2 > t_1$.
- $h^+$: A plausible hypothesis that explains the two observations $O_1$ and $O_2$.
- $h^-$: An implausible (or less plausible) hypothesis for observations $O_1$ and $O_2$.

Given the observations and a pair of hypotheses, the $\alpha$NLI task is to select the most plausible explanation (hypothesis).

**Abductive Natural Language Generation** $\alpha$NLG is the task of generating a valid hypothesis $h^+$ given the two observations $O_1$ and $O_2$. Formally, the task requires to maximize $P(h^+ | O_1, O_2)$. 

---

2 Pronounced as alpha-NLI

3 ART: Abductive Reasoning in narrative Text.
3 MODELS FOR ABDUCTIVE COMMONSENSE REASONING

3.1 ABDUCTIVE NATURAL LANGUAGE INFERENCE

A Probabilistic Framework for αNLI: A distinct feature of the αNLI task is that it requires jointly considering all available observations and their commonsense implications, to identify the correct hypothesis. Formally, the αNLI task is to select the hypothesis \( h^* \) that is most probable given the observations.

\[
h^* = \arg \max_{h^i} P(h^i|O_1, O_2)
\]  

(1)

Rewriting the objective using Bayes Rule, we have:

\[
P(h^i|O_1, O_2) \propto P(O_2|h^i, O_1)P(h^i|O_1)
\]  

(2)

We formulate a set of probabilistic models for αNLI that make various independence assumptions on Equation 2 – starting from a simple baseline that ignores the observations entirely, and building up to a fully joint model. These models are depicted as Bayesian Networks in Figure 2.

![Bayesian Networks](image)

Figure 2: Illustration of the graphical models described in the probabilistic framework. The “Fully Connected” model can, in theory, combine information from both available observations.

Hypothesis Only: Our simplest model makes the strong assumption that the hypothesis is entirely independent of both observations, i.e. \((H \perp O_1, O_2)\), in which case we simply aim to maximize the marginal \( P(H) \).

First (or Second) Observation Only: Our next two models make weaker assumptions: that the hypothesis depends on only one of the first \( O_1 \) or second \( O_2 \) observation.

Linear Chain: Our next model uses both observations, but considers each observation’s influence on the hypothesis *independently*, i.e. it does not combine information across the observations. Formally, the model assumes that the three variables \((O_1, H, O_2)\) form a linear Markov chain, where the second observation is conditionally independent of the first, given the hypothesis (i.e. \((O_1 \perp O_2|H)\)). Under this assumption, we aim to maximize a somewhat simpler objective than Equation 2:

\[
h^* = \arg \max_{h^i} P(h^i|O_1) \cdot P(O_2|h^i)\text{ where } (O_1 \perp O_2|h^i)
\]  

(3)

Fully Connected: Finally, our most sophisticated model jointly models all three random variables as in Equation 2, and can in principle combine information across both observations to choose the correct hypothesis.

To help illustrate the subtle distinction between how the Linear Chain and Fully Connected models consider both observations, consider the following example. Let observation \( O_1 \): “Carl went to the store desperately searching for flour tortillas for a recipe.” and \( O_2 \): “Carl left the store very frustrated.”. Then consider two distinct hypotheses, an incorrect \( h^1 \): “The cashier was rude” and the correct \( h^2 \): “The store had corn tortillas, but not flour ones.”. For this example, a Linear Chain model could arrive at the wrong answer, because it reasons about the observations separately—taking \( O_1 \) in isolation, both \( h^1 \) and \( h^2 \) seem plausible next events, albeit each \( a \ priori \) unlikely. And for \( O_2 \) in isolation—i.e. in the absence of \( O_1 \), as for a randomly drawn shopper—the \( h^1 \) explanation of a rude cashier seems a much more plausible explanation of Carl’s frustration than are the details of the store’s tortilla selection. Combining these two separate factors leads the Linear Chain to select \( h^1 \) as the more plausible explanation. It is only by reasoning about Carl’s goal in \( O_1 \) *jointly* with
1. Plausible Hypothesis Options: We presented collection and research in this direction. publicly release templates and the full set of instructions for all crowdsourcing tasks to facilitate future data.

nine embeddings, each associated with one commonsense inference type. to enforce the conditional independence. Specifically, we learn a discriminative classifier:

\[ P_{\text{Linear Chain}}(h|o_1, o_2) \propto e^{\phi(o_1, h) + \phi'(h, o_2)} \]

where \( \phi \) and \( \phi' \) are neural networks that produce scalar values.

3.2 Abductive Natural Language Generation

Given \( h^+ = \{ w^1, \ldots, w^h \} \), \( O_1 = \{ o_1^1, \ldots, o_1^m \} \) and \( O_2 = \{ o_2^1, \ldots, o_2^2 \} \) as sequences of tokens, the \( \alpha \)NLP task can be modeled as \( P(h^+ | O_1, O_2) = \prod P(w^h_i | w^h_{<i}, o_1^1, \ldots, o_1^m, o_2^1, \ldots, o_2^2) \). Option-

ally, the model can also be conditioned on background knowledge \( K \). Parameterized models can then be trained to minimize the negative log-likelihood over instances in \( \mathcal{ARf} \):

\[ \mathcal{L} = -\sum_{i=1}^{N} \log P(w^h_i | w^h_{<i}, o_1^1, \ldots, o_1^m, o_2^1, \ldots, o_2^2, K) \] (4)

4 \( \mathcal{ARf} \) Dataset: Abductive Reasoning in Narrative Text

\( \mathcal{ARf} \) is the first large-scale benchmark dataset for studying abductive reasoning in narrative texts. It consists of \( \sim 20K \) narrative contexts (pairs of observations \( \langle O_1, O_2 \rangle \)) with over 200K explanatory hypotheses. Table 5 in the Appendix summarizes corpus-level statistics of the \( \mathcal{ARf} \) dataset. Figure 4 shows some illustrative examples from \( \mathcal{ARf} \) (dev split). The best model based on BERT fails to correctly predict the first two dev examples.

Collecting Observations: The pairs \( O_1, O_2 \) in \( \mathcal{ARf} \) are drawn from the ROCStories dataset (Mostafazadeh et al., 2016). ROCStories is a large collection of short, manually curated five-

sentence stories. It was designed to have a clear beginning and ending for each story, which naturally map to the first \( (O_1) \) and second \( (O_2) \) observations in \( \mathcal{ARf} \).

Collecting Hypotheses Options: We crowdsourced the plausible and implausible hypotheses options on Amazon Mechanical Turk (AMT) in two separate tasks:

1. Plausible Hypothesis Options: We presented \( O_1 \) and \( O_2 \) as narrative context to crowdworkers who were prompted to fill in “What happened in-between?” in natural language. The design of
1. The task motivates the use of abductive reasoning to hypothesize likely explanations for the two given observations.

2. Implausible Hypothesis Options: In this task, we presented workers with observations $O_1, O_2$ and one plausible hypothesis option $h^+ \in H^+$ collected from the previous task. Crowdworkers were instructed to make minimal edits (up to 5 words) to a given $h^+$ to create implausible hypothesis variations for each plausible hypothesis.

A significant challenge in creating datasets is avoiding annotation artifacts – unintentional patterns in the data that leak information about the target label – that several recent studies (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018) have reported on crowdsourced datasets. To tackle this challenge, we collect multiple plausible and implausible hypotheses for each $⟨O_1, O_2⟩$ pair (as described above) and then apply an adversarial filtering algorithm to retain one challenging pair of hypotheses that are hard to distinguish between. We describe our algorithm in detail in Appendix A.5. While our final dataset uses BERT as the adversary, preliminary experiments that used GPT as an adversary resulted in similar drops in performance of all models, including all BERT variants. We compare the results of the two adversaries in Table 1.

5 Experiments and Results

We now present our evaluation of finetuned state-of-the-art pre-trained language models on the ART dataset, and several other baseline systems for both $\alpha$NLI and $\alpha$NLG. Since $\alpha$NLI is framed as a binary classification problem, we choose accuracy as our primary metric. For $\alpha$NLG, we report performance on automated metrics such as BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavie, 2005) and also report human evaluation results.

5.1 Abductive Natural Language Inference

Despite strong performance on several other NLP benchmark datasets, the best baseline model based on BERT achieves an accuracy of just 68.9% on ART compared to human performance of 91.4%. The large gap between human performance and that of the best system provides significant scope for development of more sophisticated abductive reasoning models. Our experiments show that introducing the additional independence assumptions described in Section 3.1 over the fully connected model tends to degrade system performance (see Table 1) in general.

Human Performance We compute human performance using AMT. Each instance (two observations and two hypothesis choices) is shown to three workers who were prompted to choose the more plausible hypothesis choice. We compute majority vote on the labels assigned which leads to a human accuracy of 91.4% on the ART test set.

Additional crowdsourcing details in the Appendix A.1
Baselines We include baselines that rely on simple features to verify that ART is not trivially solvable due to noticeable annotation artifacts, observed in several crowdsourced datasets. The accuracies of all simple baselines are close to chance-performance on the task – indicating that the dataset is free of simple annotation artifacts.

A model for the related but distinct task of entailment NLI (e.g. SNLI) forms a natural baseline for ART. We re-train the ESIM+ELMo (Chen et al., 2017; Peters et al., 2018) model as its performance on entailment NLI (88.9%) is close to state-of-the-art models (excluding pre-trained language models). This model only achieves an accuracy of 58.8%, highlighting that performing well on ART requires models to go far beyond the linguistic notion of entailment.

<table>
<thead>
<tr>
<th>Model</th>
<th>GPT AF Acc. (%)</th>
<th>ART Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random (2-way choice)</td>
<td>50.1</td>
<td>50.4</td>
</tr>
<tr>
<td>Majority (from dev set)</td>
<td>50.1</td>
<td>50.8</td>
</tr>
<tr>
<td>Infersent (Conneau et al., 2017)</td>
<td>50.9</td>
<td>50.8</td>
</tr>
<tr>
<td>ESIM+ELMo (Chen et al., 2017)</td>
<td>58.2</td>
<td>58.8</td>
</tr>
</tbody>
</table>

Finetuning Pre-trained LMs

<table>
<thead>
<tr>
<th>Model</th>
<th>GPT AF Acc. (%)</th>
<th>ART Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-ft</td>
<td>52.6 (0.9)</td>
<td>63.1 (0.5)</td>
</tr>
<tr>
<td>BERT-ft [h Only]</td>
<td>55.9 (0.7)</td>
<td>59.5 (0.2)</td>
</tr>
<tr>
<td>BERT-ft [O₁ Only]</td>
<td>63.9 (0.8)</td>
<td>63.5 (0.7)</td>
</tr>
<tr>
<td>BERT-ft [O₂ Only]</td>
<td>68.1 (0.6)</td>
<td>66.6 (0.2)</td>
</tr>
<tr>
<td>BERT-ft [Linear Chain]</td>
<td>65.3 (1.4)</td>
<td>68.9 (0.5)</td>
</tr>
<tr>
<td>BERT-ft [Fully Connected]</td>
<td>72.0 (0.5)</td>
<td>68.6 (0.5)</td>
</tr>
</tbody>
</table>

Human Performance

- 91.4

Table 1: Performance of baselines and finetuned-LM approaches on the test set of ART. Test accuracy is reported as the mean of five models trained with random seeds, with the standard deviation in parenthesis.

Learning Curve and Dataset Size

While there is enough scope for considerably scaling up the dataset based on ROCStories, the learning curve in Figure 5 shows that the performance of the best model plateaus after ∼10,000 instances. In addition, there is still a wide gap (∼23%) between the performance of the best model and human performance.

GPT Adversary Table 1 also includes results of our experiments where GPT was used as the adversary. Notably, in this case, adversarially filtering the dataset brings down GPT performance under 53%. On the other hand, the best BERT model, that encodes the fully connected bayesian network performs significantly better than the BERT model that encodes the linear chain assumptions – 72% compared to 65%. Therefore, we use the BERT fully connected model as the adversary in ART. The gap between the linear chain and fully connected BERT models diminishes when BERT is used as an adversary – in spite of being a more powerful model – which indicates that adversarial filtering disproportionately impacts the model used as the adversary. However, the dataset also becomes more difficult for the other models that were not used as adversaries. For example, before any filtering, BERT scores 88% and OpenGPT gets 80%, which is much higher than either model achieves in Table 1 when the other

The input format for the GPT model and BERT variants is described in the Appendix A.4.
Table 2: Performance of generative models on the test set of ART. All models except GPT2-Fixed are finetuned on ART.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>METEOR</th>
<th>ROUGE</th>
<th>CIDEr</th>
<th>BERT-Score</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT2-Fixed</td>
<td>0.0</td>
<td>9.29</td>
<td>9.99</td>
<td>3.34</td>
<td>36.69</td>
<td>-</td>
</tr>
<tr>
<td>O₁-O₂-Only</td>
<td>2.23</td>
<td>16.71</td>
<td>22.83</td>
<td>33.54</td>
<td>48.74</td>
<td>42.26</td>
</tr>
<tr>
<td>COMeT-Txt+GPT2</td>
<td>2.29</td>
<td>16.73</td>
<td>22.51</td>
<td>32.00</td>
<td>48.52</td>
<td>44.56</td>
</tr>
<tr>
<td>COMeT-Emb+GPT2</td>
<td>3.03</td>
<td>17.66</td>
<td>22.93</td>
<td>33.54</td>
<td>48.74</td>
<td>42.26</td>
</tr>
<tr>
<td>Human-written Hypotheses</td>
<td>8.25</td>
<td>26.71</td>
<td>30.40</td>
<td>53.56</td>
<td>53.30</td>
<td>96.03</td>
</tr>
</tbody>
</table>

Table 2: Performance of generative models on the test set of ART. All models except GPT2-Fixed are finetuned on ART.

model is used for filtering. This result is a reasonable indicator, albeit not a guarantee, that ART will remain challenging for new models released in the future.

5.2 ABDUCTIVE NATURAL LANGUAGE GENERATION

Generative Language Models As described in Equation 4, we train GPT2 conditioned on the tokens of the two observations $O_1$ and $O_2$. Both observations are enclosed with field-specific tags. ATOMIC (Sap et al., 2019), a repository of inferential if-then knowledge is a natural source of background commonsense required to reason about narrative contexts in ART. Yet, there is no straightforward way to include such knowledge into a neural model as ATOMIC’s nodes are not canonicalized and are represented as short phrases of text. Thus, we rely on COMeT – a transformer model trained on ATOMIC that generates nine commonsense inferences of events in natural language. Specifically, we experiment with two ways of integrating information from COMeT in GPT2: (i) as phrases, and (ii) as embeddings.

Figure 3 shows how we integrate COMeT representations. Concretely, after the input tokens are embedded by the word-embedding layer, we append eighteen (nine relations for each observation) embeddings to the sequence before passing through the layers of the Transformer architecture. This allows the model to learn each token’s representation while attending to the COMeT embeddings – effectively integrating background commonsense knowledge into a language model.

Discussion Table 2 reports results on the αNLG task. Among automatic metrics, we report BLEU-4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) and BERT-Score (Zhang et al., 2019) (with the bert-base-uncased model). We establish human performance through crowdsourcing on AMT. Crowdworkers are shown pairs of observations and a generated hypothesis and asked to label whether the hypothesis explains the given observations. The last column reports the human evaluation score. The last row reports the score of a held-out human-written hypothesis and serves as a ceiling for model performance. Human-written hypotheses are found to be correct for 96% of instances, while our best generative models, even when enhanced with background commonsense knowledge, only achieve 45% – indicating that the αNLG generation task is especially challenging for current state-of-the-art text generators.

8Please see Appendix A.6 for a full list of the nine relations.
9We describe the format of input for each model in Appendix A.7.
6 ANALYSIS

6.1 αNLI

Commonsense reasoning categories We investigate the categories of commonsense-based abductive reasoning that are challenging for current systems and the ones where the best model outperforms. While there have been previous attempts to categorize commonsense knowledge required for entailment (LoBue & Yates, 2011; Clark et al., 2007), crowdsourcing this task at scale with high fidelity and high agreement across annotators remains challenging. Instead, we aim to probe the model with soft categories identified by matching lists of category-specific keywords to the hypothesis choices.

<table>
<thead>
<tr>
<th>Category</th>
<th>Human Accuracy</th>
<th>BERT-ft Accuracy</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>All (1,000)</td>
<td>91.4</td>
<td>68.8</td>
<td>22.6</td>
</tr>
<tr>
<td>Numerical (44)</td>
<td>88.6</td>
<td>56.8</td>
<td>21.8</td>
</tr>
<tr>
<td>Spatial (130)</td>
<td>91.5</td>
<td>65.4</td>
<td>26.1</td>
</tr>
<tr>
<td>Emotional (84)</td>
<td>86.9</td>
<td>72.6</td>
<td>14.3</td>
</tr>
</tbody>
</table>

Table 3: BERT’s performance and human evaluation on categories for 1,000 instances from the test set, based on commonsense reasoning domains (Numerical, Spatial, Emotional). The number in parenthesis indicates the size of the category.

Table 3 shows the accuracy of the best model (BERT-ft) across various categories of commonsense knowledge. BERT-ft significantly underperforms on instances involving Numerical (56.8%) and Spatial (65.4%) commonsense. These two categories include reasoning about numerical quantities and the spatial location of agents and objects, and highlight some of the limitations of the language models. In contrast, it significantly overperforms on the Emotional category (72.6%) where the hypotheses exhibit strong textual cues about emotions and sentiments.

Implausible transitions A model for an instance of the ART dataset should discard implausible hypotheses in the context of the two given observations. In narrative contexts, there are three main reasons for an implausible hypothesis to be labeled as such:

1. $O_1 \rightarrow h^-: h^-$ is unlikely to follow after the first observation $O_1$.
2. $h^- \rightarrow O_2$: $h^-$ is plausible after $O_1$ but unlikely to precede the second observation $O_2$.
3. Plausible: $(O_1, h^-, O_2)$ is a coherent narrative and forms a plausible alternative, but it is less plausible than $(O_1, h^+, O_2)$.

We analyze the prevalence of each of these reasons in ART. We design a crowdsourcing task in which we show the implausible option along with the narrative context $(O_1, O_2)$ and get labels for which transition $(O_1 \rightarrow h^-, h^- \rightarrow O_2$ or neither) in the narrative chain is broken. Table 4 shows the proportion of each category from a subset of 1,000 instances from the test set. While $h^- \rightarrow O_2$ accounts for almost half of the implausible transitions in ART, all three categories are substantially present in the dataset. BERT performance on each of these categories indicates that the model finds it particularly hard when the narrative created by the incorrect hypothesis is plausible, but less plausible than the correct hypothesis. On that subset of the test set, the fully connected model performs better than the linear chain model where it is important to consider both observations jointly to arrive at the more likely hypothesis.

<table>
<thead>
<tr>
<th>Story Transition</th>
<th>% of Dataset</th>
<th>BERT-ft Fully Connected Acc. (%)</th>
<th>BERT-ft Linear Chain Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_1 \rightarrow h^-$</td>
<td>32.5</td>
<td>73.6</td>
<td>71.6</td>
</tr>
<tr>
<td>$h^- \rightarrow O_2$</td>
<td>45.3</td>
<td>69.0</td>
<td>70.5</td>
</tr>
<tr>
<td>Plausible</td>
<td>22.2</td>
<td>62.5</td>
<td>58.5</td>
</tr>
<tr>
<td>All (1,000)</td>
<td>100.0</td>
<td>69.1</td>
<td>68.2</td>
</tr>
</tbody>
</table>

Table 4: Fraction of dataset for which a particular transition in the story is broken for the negative hypothesis, for 1,000 random instances from the test set.

6.2 αNLG

Figure 7 shows some examples of generations from the trained models compared to human-written generations. The example on the left is an example of an instance that only humans could get correct, while for the one on the right, COMET-Emb+GPT2 also generates the correct explanation for the observations.
Larry’s yard was covered in dead leaves.

Larry decided to give up for the day and went back inside.

Junior is the name of our 20+ year old turtle.

Junior is still going strong.

Larry wondered what he could do with the leaves.

He spent hours trying to clean the yard.

Larry threw the leaves out.

Junior will have surgery to heal and her internal organs are broken.

We took Junior to the vet to check on him.

Generated Hypotheses

<table>
<thead>
<tr>
<th>Model</th>
<th>Hypothesis</th>
<th>Validated</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPT2</td>
<td>actually, that turtle can’t bite you</td>
<td></td>
</tr>
<tr>
<td>COMeT-Text+GPT2</td>
<td>Junior will have surgery to heal and her internal organs are broken</td>
<td></td>
</tr>
<tr>
<td>COMeT-Emb+GPT2</td>
<td>Junior has been swimming in the pool with her friends</td>
<td></td>
</tr>
<tr>
<td>Human-written</td>
<td>We took Junior to the vet to check on him</td>
<td></td>
</tr>
</tbody>
</table>

Figure 7: Examples of generated hypotheses from different models and human-written hypothesis for 2 instances from ART.

7 Related Work

Cloze-Style Task vs. Abductive Reasoning
Since abduction is fundamentally concerned with plausible chains of cause-and-effect, our work draws inspiration from previous works that deal with narratives such as script learning (Schank & Abelson, 1975) and the narrative cloze test (Chambers & Jurafsky, 2009; Jans et al., 2012; Pichotta & Mooney, 2014; Rudinger et al., 2015). Rather than learning prototypical scripts or narrative chains, we instead reason about the most plausible events conditioned on observations. We make use of the ROCStories dataset (Mostafazadeh et al., 2016), which was specifically designed for the narrative cloze task. But, instead of reasoning about plausible event sequences, our task requires reasoning about plausible explanations for narrative omissions.

Entailment vs. Abductive Reasoning The formulation of αNLI is closely related to entailment NLI, but there are two critical distinctions that make abductive reasoning uniquely challenging. First, abduction requires reasoning about commonsense implications of observations (e.g., if we observe that the “grass is wet”, a likely hypothesis is that “it rained earlier”) which go beyond the linguistic notion of entailment (also noted by Josephson (2000)). Second, abduction requires non-monotonic reasoning about a set of commonsense implications collectively, to check the potential contradictions against multiple observations and to compare the level of plausibility of different hypotheses. This makes abductive reasoning distinctly challenging compared to other forms of reasoning such as induction and deduction (Shank, 1998). Perhaps more importantly, abduction is closely related to the kind of reasoning humans perform in everyday situations, where information is incomplete and definite inferences cannot be made.

Generative Language Modeling Recent advancements in the development of large-scale pre-trained language models (Radford, 2018; Devlin et al., 2018; Radford et al., 2019) have improved the quality and coherence of generated language. Although these models have shown to generate reasonably coherent text when condition on a sequence of text, our experiments highlight the limitations of these models to 1) generate language non-monotonically and 2) adhere to commonsense knowledge. We attempt to overcome these limitations with the incorporation of a generative commonsense model during hypothesis generation.

Related Datasets Our new resource A RT complements ongoing efforts in building resources for natural language inference (Dagan et al., 2006; MacCartney & Manning, 2009; Bowman et al., 2015; Williams et al., 2018a; Camburu et al., 2018). Existing datasets have mostly focused on textual entailment in a deductive reasoning set-up (Bowman et al., 2015; Williams et al., 2018a) and making inferences about plausible events (Maslan et al., 2015; Zhang et al., 2017). In their typical setting, these datasets require a system to deduce the logically entailed consequences of a given premise. In contrast, the nature of abduction requires the use of commonsense reasoning capabilities, with less focus on lexical entailment. While abductive reasoning has been applied
to entailment datasets (Raina et al., 2005), they have been applied in a logical theorem-proving framework as an intermediate step to perform textual entailment – a fundamentally different task than NLI.

8 CONCLUSION

We present the first study that investigates the viability of language-based abductive reasoning. We conceptualize and introduce Abductive Natural Language Inference (αNLI) – a novel task focused on abductive reasoning in narrative contexts. The task is formulated as a multiple-choice question-answering problem. We also introduce Abductive Natural Language Generation (αNLG) – a novel task that requires machines to generate plausible hypotheses for given observations. To support these tasks, we create and introduce a new challenge dataset, ART, which consists of 20,000 commonsense narratives accompanied with over 200,000 explanatory hypotheses. In our experiments, we establish comprehensive baseline performance on this new task based on state-of-the-art NLI and language models, which leads to 68.9% accuracy with a considerable gap with human performance (91.4%). The αNLG task is significantly harder – while humans can write a valid explanation 96% of times, the best generator models can only achieve 45%. Our analysis leads to new insights into the types of reasoning that deep pre-trained language models fail to perform – despite their strong performance on the closely related but different task of entailment NLI – pointing to interesting avenues for future research. We hope that ART will serve as a challenging benchmark for future research in language-based abductive reasoning and the αNLI and αNLG tasks will encourage representation learning that enables complex reasoning capabilities in AI systems.

REFERENCES


A APPENDICES

A.1 DATA COLLECTION DETAILS

We describe the crowdsourcing details of our data collection method.

**Task 1 - Plausible Hypothesis Options** In this task, participants were presented an incomplete three-part story, which consisted of the first observation \((O_1)\) and the second observation \((O_2)\) of the story. They were then asked to complete the story by writing a probable middle sentence that explains why the second observation should follow after the first one. We instructed participants to make sure that the plausible middle sentence (1) is short (fewer than 10 words) and (2) simple as if narrating to a child, (3) avoids introducing any extraneous information, and (4) uses names instead of pronouns (e.g., he/she) wherever possible.

All participants were required to meet the following qualification requirements: (1) their location is in the US, (2) HIT approval rate is greater than 95\(\%\), and (3) Number of HITs approved is greater than 5,000. The reward of this task was set to be $0.07 per question ($14/hour in average), and each HIT was assigned to five different workers (i.e., 5-way redundancy).

**Task 2 - Implausible Hypothesis Options** In this task, participants were presented a three-part story, which consisted of the first observation \((O_1)\), a middle sentence \((h^+)\) collected in Task 1, and the second observation \((O_2)\) of the story. They were then asked to rewrite the middle sentence \((h^+)\) with minimal changes, so that the story becomes unlikely, implausible or inconsistent \((h^-)\). We asked participants to add or remove at most four words to \(h^+\), while ensuring that the new middle sentence is grammatical. In addition, we asked them to stick to the context in the given story. For example, if the story talks about “doctors”, they are welcome to talk about “health” or “diagnosis”, but not mention “aliens”. Finally, we also asked workers to verify if the given middle \((h^+)\) makes a plausible story, in order to confirm the plausibility of \(h^+\) collected in Task 1.

With respect to this task’s qualification, participants were required to fulfill the following requirements: (1) their location is the US or Canada, (2) HIT approval rate is greater than or equal to 99\(\%\), and (3) number of HITs approved is greater than or equal to 10,000. Participants were paid $0.1 per question ($14/hour in average), and each HIT was assigned to three different participants (i.e., 3-way redundancy).

**Task 3 - \(\alpha\)NLI Human Performance** Human performance was evaluated by asking participants to answer the \(\alpha\)NLI questions. Given a narrative context \((O_1, O_2)\) and two hypotheses, they were asked to choose the more plausible hypothesis. They were also allowed to choose “None of the above” when neither hypothesis was deemed plausible.

We asked each question to seven participants with the following qualification requirements: (1) their location is either in the US, UK, or Canada, (2) HIT approval rate is greater than 98\(\%\), and (3) number of HITs approved is greater than or equal to 10,000. The reward was set to $0.05 per HIT. We took the majority vote among the seven participants for every question to compute human performance.

A.2 ART DATA STATISTICS

Table 5 shows some statistics of the ART dataset.

A.3 FINE-TUNING BERT

We fine-tuned the BERT model using a grid search with the following set of hyper-parameters:

- batch size: \{3, 4, 8\}
- number of epochs: \{3, 4, 10\}
- learning rate: \{1e-5, 2e-5, 3e-5, 5e-5\}

The warmup proportion was set to 0.2, and cross-entropy was used for computing the loss. The best performance was obtained with a batch size of 4, learning rate of 5e-5, and number of epochs equal to 10. Table 6 describes the input format for GPT and BERT (and its variants).
Table 5: Some statistics summarizing the ART dataset. The train set includes all plausible and implausible hypotheses collected via crowdsourcing, while the dev and test sets include the hypotheses selected through the Adversarial Filtering algorithm.

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total unique occurrences</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contexts ( (O_1, O_2) )</td>
<td>14,196</td>
<td>2,010</td>
<td>4,056</td>
</tr>
<tr>
<td>Plausible hyp. ( h^+ )</td>
<td>57,803</td>
<td>2,010</td>
<td>4,056</td>
</tr>
<tr>
<td>Implausible hyp. ( h^- )</td>
<td>124,692</td>
<td>2,010</td>
<td>4,056</td>
</tr>
<tr>
<td><strong>Avg. size per context</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plausible hyp. ( h^+ )</td>
<td>4.1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Implausible hyp. ( h^- )</td>
<td>8.8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Avg. word length</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plausible hyp. ( h^+ )</td>
<td>8.4</td>
<td>8.6</td>
<td>8.5</td>
</tr>
<tr>
<td>Implausible hyp. ( h^- )</td>
<td>8.2</td>
<td>8.6</td>
<td>8.4</td>
</tr>
<tr>
<td>First observation ( O_1 )</td>
<td>8.1</td>
<td>8.1</td>
<td>8.0</td>
</tr>
<tr>
<td>Second observation ( O_2 )</td>
<td>9.3</td>
<td>9.4</td>
<td>9.2</td>
</tr>
</tbody>
</table>

A.4 Baselines

The SVM classifier is trained on simple features like word length, overlap and sentiment features to select one of the two hypothesis choices. The bag-of-words baseline computes the average of GloVe embeddings for words in each sentence to form sentence embeddings. The sentence embeddings in a story (two observations and a hypothesis option) are concatenated and passed through fully-connected layers to produce a score for each hypothesis. The accuracies of both baselines are close to 50% (SVM: 50.6; BOW: 50.5).

Specifically, we train an SVM classifier and a bag-of-words model using GLoVE embeddings. Both models achieve accuracies close to 50%. An InferSent (Conneau et al., 2017) baseline that uses sentences embedded by max-pooling over Bi-LSTM token representations achieves only 50.8% accuracy.

Table 6: Input formats for GPT and BERT fine-tuning.

<table>
<thead>
<tr>
<th>Model</th>
<th>Input Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT</td>
<td>([\text{START}] \ O_1 + h^+ \ [\text{SEP}] \ O_2 \ [\text{SEP}])</td>
</tr>
<tr>
<td>BERT-ft [Hypothesis Only]</td>
<td>([\text{CLS}] \ h^+ \ [\text{SEP}])</td>
</tr>
<tr>
<td>BERT-ft [First Observation Only]</td>
<td>([\text{CLS}] \ O_1 \ [\text{SEP}] \ h^+ \ [\text{SEP}])</td>
</tr>
<tr>
<td>BERT-ft [Second Observation Only]</td>
<td>([\text{CLS}] \ h^+ \ [\text{SEP}] \ O_2 \ [\text{SEP}])</td>
</tr>
<tr>
<td>BERT-ft [Linear Chain]</td>
<td>([\text{CLS}] \ O_1 \ [\text{SEP}] \ h^+ \ [\text{SEP}] ; [\text{CLS}] \ h^+ \ [\text{SEP}] \ O_2 \ [\text{SEP}])</td>
</tr>
<tr>
<td>BERT-ft [Fully Connected]</td>
<td>([\text{CLS}] \ O_1 + O_2 \ [\text{SEP}] \ h^+ \ [\text{SEP}])</td>
</tr>
</tbody>
</table>

A.5 Adversarial Filtering of Hypotheses Choices

Given an observation pair and sets of plausible and implausible hypotheses \( (O_1, O_2, H^+, H^-) \), our adversarial filtering algorithm selects one plausible and one implausible hypothesis \( (O_1, O_2, h^+, h^-) \) such that \( h^+ \) and \( h^- \) are hard to distinguish between. We make three key improvements over the previously proposed Adversarial Filtering (AF) approach in Zellers et al. (2018). First, Instead of a single positive sample, we exploit a pool \( H^+ \) of positive samples to choose from (i.e. plausible hypotheses). Second, Instead of machine generated distractors, the pool \( H^- \) of negative samples (i.e. implausible hypotheses) is human-generated. Thus, the distractors share stylistic features of the positive samples as well as that of the context (i.e. observations \( O_1 \) and \( O_2 \)) – making the negative samples harder to distinguish from positive samples. Finally, We use BERT (Devlin et al., 2018) as the adversary and introduce a temperature parameter that controls the maximum number of instances
that can be modified in each iteration of AF. In later iterations, fewer instances get modified resulting in a smoother convergence of the AF algorithm (described in more detail below).

Algorithm 1 provides a formal description of our approach. In each iteration $i$, we train an adversarial model $M_i$ on a random subset $T_i$ of the data and update the validation set $V_i$ to make it more challenging for $M_i$. For a pair $(h^+_k, h^-_k)$ of plausible and implausible hypotheses for an instance $k$, we denote $\delta = \Delta M_i(h^+_k, h^-_k)$ the difference in the model evaluation of $h^+_k$ and $h^-_k$. A positive value of $\delta$ indicates that the model $M_i$ favors the plausible hypothesis $h^+_k$ over the implausible one $h^-_k$.

With probability $t_i$, we update instance $k$ that $M_i$ gets correct with a pair $(h^+_k, h^-_k) \in H^+_k \times H^-_k$ of hypotheses that reduces the value of $\delta$, where $H^+_k$ (resp. $H^-_k$) is the pool of plausible (resp. implausible) hypotheses for instance $k$.

We ran AF for 50 iterations and the temperature $t_i$ follows a sigmoid function, parameterized by the iteration number, between $t_s = 1.0$ and $t_e = 0.2$. Our final dataset, ART, is generated using BERT as the adversary in Algorithm 1.

**Algorithm 1: Dual Adversarial Filtering**

**input**: dataset $D_0$, plausible & implausible hypothesis sets ($H^+, H^-$), number of iterations $n$, initial & final temperatures ($t_s, t_e$)

**output**: dataset $D_n$

1. for iteration $i : 0, n-1$ do
   2. $t_i = t_e + \frac{t_e - t_s}{1 + e^{-(i-3)/n}}$
   3. Randomly partition $D_i$ into $(T_i, V_i)$.
   4. Train model $M_i$ on $T_i$.
   5. $S_i = \emptyset$, the selected hypotheses for $V_i$.
   6. for $(h^+_k, h^-_k) \in V_i$ do
      7. Pick $r$ uniformly at random in $[0, 1]$.
      8. if $r > t_i$ or $\Delta M_i(h^+_k, h^-_k) < 0$ then
         9. Add $(h^+_k, h^-_k)$ to $S_i$.
      10. else
         11. Pick $(h^+, h^-) \in H^+_k \times H^-_k$ s.t. $\Delta M_i(h^+, h^-) < \Delta M_i(h^+_k, h^-_k)$
         12. Add $(h^+, h^-)$ to $S_i$.
   13. end
   14. $D_{i+1} = T_i \cup S_i$
15. end

A.6 ATOMIC RELATIONS

ATOMIC (Sap et al., 2019) represents commonsense knowledge as a graph with events as nodes and the following nine relations as edges:

1. xIntent: Why does X cause an event?
2. xNeed: What does X need to do before the event?
3. xAttr: How would X be described?
4. xEffect: What effects does the event have on X?
5. xWant: What would X likely want to do after the event?
6. xReaction: How does X feel after the event?
7. oReact: How do others feel after the event?
8. oWant: What would others likely want to do after the event?
9. oEffect: What effects does the event have on others?

A.7 GENERATION MODELS INPUT FORMAT
### Table 7: Input format used to training and generated text from various GPT2 based models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Input Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT2-Fixed</td>
<td>$w_1 \ldots w_n, w_1^2 \ldots w_n^2$ Because,</td>
</tr>
<tr>
<td>$O_1$-$O_2$-Only</td>
<td>$(o1)w_1 \ldots w_n^2(o2)w_1^2 \ldots w_n^2(o2)$</td>
</tr>
<tr>
<td>COMeT-Txt+GPT2</td>
<td>$(p_1^1)T_1^1 \ldots T_9^1(p_2^1)T_1^2 \ldots T_9^2(p_2^2)w_1 \ldots w_n^1(o1)w_1^2 \ldots w_n^2(o2)$</td>
</tr>
<tr>
<td>COMeT-Emb+GPT2</td>
<td>$c_1 \ldots c_9, c_1^2 \ldots c_9^2(o1)w_1 \ldots w_n^1(o1)w_1^2 \ldots w_n^2(o2)$</td>
</tr>
</tbody>
</table>

$c_i^j$ refers to the COMeT embeddings obtained using a separate transformer model for relation $i$ and observation $j$. Similarly, $T_i^j$ is the textual phrase for relation $i$, observation $j$. Where appropriate, field specific start and end-tags are added to the sequence of inputs.