Improving Commonsense Causal Reasoning by Adversarial Training and Data Augmentation

Ieva Staliūnaitė, Philip John Gorinski, Ignacio Iacobacci
Huawei Noah's Ark Lab, London, United Kingdom
{ieva.staliunaite | philip.john.gorinski | ignacio.iacobacci}@huawei.com

Abstract

Determining the plausibility of causal relations between clauses is a commonsense reasoning task that requires complex inference ability. The general approach to this task is to train a large pretrained language model on a specific dataset. However, the available training data for the task is often scarce, which leads to instability of model training or reliance on the shallow features of the dataset. This paper presents a number of techniques for making models more robust in the domain of causal reasoning. Firstly, we perform adversarial training by generating perturbed inputs through synonym substitution. Secondly, based on a linguistic theory of discourse connectives, we perform data augmentation using a discourse parser for detecting causally linked clauses in large text, and generating distractors with a generative language model. Both methods boost model performance on the Choice of Plausible Alternatives (COPA) dataset, as well as on a Balanced COPA dataset, which is a modified version of the original data that has been developed to avoid superficial cues, leading to a more challenging benchmark. We show a statistically significant improvement in performance and robustness on both datasets, even with only a small number of additionally generated data points.

Introduction

Within the discourse discipline in linguistic research, causal relations are classified as *contingency relations*, which in turn are a subset of coherence relations. Coherence relations are temporal, comparison, expansion and contingency relations between clauses, which can be explicitly marked with certain discourse connectives (Asr and Demberg 2013). Examples of the expressions of the four types of coherence relations include "*A then/before/while B*", "*A even though/but/however B*", "*A and/moreover/or B*" and "*A because/therefore/if B*", respectively. In this work we are specifically interested in the causal relations, which can be split into backward (e.g. expressed as "*A because B*") and forward (e.g. expressed as "*A therefore B*") ones, differing with regard to whether A or B refers to the cause of the event expressed by the other clause. It is interesting that causal connectives such as "*because*" convey a lot of information needed for inferring the causal relation between clauses. That is, the discourse connective "*because*" disambiguates the causal relation between the clauses more than most other connectives (Asr and Demberg 2013).

In Natural Language Processing (NLP), the task of causal inference is to determine the presence of a causal link *without* the cue of a discourse connective. Commonsense causal reasoning is a complex task, which requires not only linguistic parsing and logical inference but also world knowledge of causal links between events. For example, if a model was to expertly determine which of the two alternatives is more likely to be caused by the premise below, it would have to access knowledge of the causal link between the ripening and edibility of bananas.

Premise: The bananas ripened.
Alternative1: We squeezed them.
Alternative2: We ate them.
Question: Effect
Label: Alternative2

From Roemmele, Bejan, and Gordon (2011)

The most commonly used benchmark task for evaluation of commonsense reasoning models is the the Choice of Plausible Alternatives (Roemmele, Bejan, and Gordon 2011, COPA). In this task the goal is to determine which of the two given alternatives is the *true* choice, i.e. which alternative is causally linked to a given premise, and which one is a *distractor*. This is an easier task than determining the precise coherence relation between two clauses. The task is relatively easy for humans, as shown by the nearly perfect annotator agreement for this task (Cohen's kappa = 0.965). However, the COPA dataset is very small (1000 items in total), and high model performance on it is not stable and often relies on biases in the data (Kavumba et al. 2019). Consequently, Kavumba et al. (2019) introduce a Balanced COPA dataset by manually adjusting items from COPA to remove the superficial features, such as patterns of determiner use, which are generally shown to be exploited by models for solving the task. Data points are generated by mirroring each example in a way that it captures a similar relation, yet with the biasing feature appearing in the alternative which does not manifest that feature in the original data. That way the biasing feature appears in the *true* choice in one case and in the *distractor* in the other case. For example, Kavumba et al.

(2019) show that models rely on indefinite determiners like "*a*" when classifying instances, as their distribution is not uniform between the classes. A mirrored example is shown here:

Original premise: The woman hummed to herself.
Original alternative1 (true): She was in **a** good mood.
Original alternative2 (distractor): She was nervous.
Mirrored premise: The woman trembled.
Mirrored alternative1 (distractor): She was in **a** good mood.
Mirrored alternative2 (true): She was nervous.

From Kavumba et al. (2019)

This leads to a dataset which is double the size of the original COPA dev set and more challenging to solve than the original COPA.

As briefly discussed by Gordon, Bejan, and Sagae (2011), the research on causal inference tasks can be split into approaches that focus on the *depth* or *breadth* of the solution. Deep approaches aim to gain more information about a particular input, for example by using knowledge graphs to learn more about the entities and events mentioned in the input or applying formal logic to deterministically find the relation that is sought for, in order to classify that input correctly (Furbach, Gordon, and Schon 2015; Furbach and Schon 2016; Blass and Forbus 2017; Siebert, Schon, and Stolzenburg 2019; Goodwin and Demner-Fushman 2019). On the other hand, the approaches of the broad type attempt to cover a wider range of instances in their method, for example by determining the types of syntactic or semantic features that are generally used in expressing a particular relation (Gordon, Bejan, and Sagae 2011; Goodwin et al. 2012; Jabeen, Gao, and Andreae 2014; Rahimtoroghi, Hernandez, and Walker 2017; Tamborrino et al. 2020; Iter et al. 2020). This paper attempts to tackle the causal reasoning task with two approaches, one of the *deep* and one of the *broad* type. That is, the first approach is adversarial training - perturbing the original inputs to produce similar but more difficult examples, which can be seen as data points in the surrounding area of the original ones. The aim of covering the area around the original datapoints is to provide the model with more semantic information about the inputs. The second approach is data augmentation by means of generating completely unseen examples, which leads to new data points further away from the original ones. The aim of covering a wider range of examples is to find general patterns of causally linked clauses.

In both methods we study to what extent model performance on the task of causal inference can be improved by augmenting the training data with linguistic information. To this end, we rely on findings of psycholinguistic research, discourse annotations, lexical semantics and language models for tackling the COPA task, using a RoBERTa model (Liu et al. 2019) as our baseline. The first approach is an application of an adversarial example generation through perturbations of the original COPA data by substituting words with their equivalent terms given the context using a semantic network, following Zang et al. (2020). The second approach consists of gathering causally linked clauses from the web with the help of a Penn Discourse TreeBank (Prasad et al. 2008, PDTB) parser (Lin, Ng, and Kan 2014) and generating the distractor alternatives by applying a technique based on discourse connectives and their denotations with the help of GPT-2 (Radford et al. 2019). Both approaches lead to varying improvements on the performance of the RoBERTa model on both the COPA and balanced COPA datasets, exhibited by higher average accuracy scores and smaller standard deviation ranges. The three main contributions in this paper can be summarized as:

1. A novel application of linguistic knowledge to the task of causal reasoning;
2. A productive augmentation method for the task of choosing plausible causal alternatives;
3. A significant improvement on the performance and robustness of a RoBERTa model on the COPA dataset.

Related Work

A wide variety of approaches have been used for tackling the task of COPA. The previously used *deep* approaches include formal linguistic methods such as the use of theorem proving on text converted to a logical form, and structured knowledge methods such as the use of knowledge graphs. The *broad* approaches include more heuristic methods such as assuming the presence of causal relations between colocated sentences in narrative text, and Neural Network approaches such as adjusting loss functions or pre-training language models with an unconventional target.

The formal linguistic representation approaches (Furbach, Gordon, and Schon 2015; Furbach and Schon 2016; Blass and Forbus 2017; Siebert, Schon, and Stolzenburg 2019) vary with regard to their focus - from tackling the problem of chaining multiple logical inferences by analogy, to merging the formal representation approach with a Neural Network model. While the theory proving approach guarantees the correct outcome given the correct representation, the representations are not easy to produce, especially for colloquial use of language.

For incorporating background knowledge, some previous research has adopted structured knowledge approaches such as knowledge graphs (Goodwin and Demner-Fushman 2019) and semantic networks (Mtarji 2019). Knowledge graphs about entities as well as causal links between event words provide information that is relevant to the type of reasoning required for commonsense causal reasoning tasks, which is assumed to be emitted in raw text based on pragmatic rules of language. Hence, as expected, the use of external world knowledge improves the performance of the prevailing neural network systems.

In addition, external linguistic knowledge sources have also been used for training models, which can be combined with other systems to inject linguistic information into the causal reasoning systems as well. For example, Goodwin et al. (2012) use syntactic dependency trees as well as annotations of temporal relations between events. In the above mentioned approaches the authors rely on assumptions such as that an event *A* is not likely causally linked to event *B* if words in *A* are not conceptually linked to words in the context of *B*, that causal relations are exhibited in certain syntactic patterns in text, and that causal relations tend to fall into a temporally bound pattern. Finally, some research uses a discourse parser trained on manually annotated causally linked pairs of clauses and sentence similarity metrics to determine the more likely causal links (Gordon, Bejan, and Sagae 2011). This approach is based on the expectation that similar sentences will stand in similar causal relations.

The latter approach differs from the aforementioned ones in that it does not annotate the words or sentences in the target clauses, but relies on other input through similarity. Similarly, a slightly different set of assumptions allows one to use only raw text or more coarse annotations in training causal reasoning systems with some fruitful outcomes as well. For example, a few papers present work based on the idea that narrative text is composed of sequences of causally linked sentences. Gordon, Bejan, and Sagae (2011), in addition to the previously discussed method, also train a model on sentences from personal blogs that appear in close proximity. The hypothesis in this method is that there is some link between the proximity of sentences in personal stories as people tend to tell them in causally linked sequences. In a very similar vein, Rahimtoroghi, Hernandez, and Walker (2017) propose learning which events are contingent on others by expecting that the sentences describing the *effects* appear *after* sentences describing the *causes* in a narrative. They show that they are able to learn some new relations that were not present in structured resources at the time. Similarly, Jabeen, Gao, and Andreae (2014) show that the order in which two events appear is relevant to finding contingency relations, as asymmetrical labels of events lead to better causal inference results than symmetrical ones.

Finally, large pre-trained Neural Network models are very prevalent in NLP tasks including causal reasoning, with the current state-of-the-art result on COPA belonging to the T-5 model finetuned for the downstream task (Raffel et al. 2019). Beside the use of out-of-the-box models, some research has proposed a different pre-training technique that is specifically aiming to improve discourse tasks. More specifically, Iter et al. (2020) use the distance between sentences as an additional training task, under the same assumption as the research discussed in the previous paragraph - namely the expectation of causal relations between subsequent sentences. They show an improvement on discourse-related tasks such as COPA among others, when a BERT model (Devlin et al. 2019) is trained with this additional objective. In addition, Shwartz et al. (2020) use a self-talk method in which they use pre-trained generative language models to produce questions and answers about entities mentioned in the target sentences in a causal reasoning task. This provides the background knowledge in a novel way, as an alternative to the use of knowledge graphs mentioned at the beginning of this section. In contrast, Tamborrino et al. (2020) rephrase the causal reasoning task into an "*A because B*" format and calculate the likelihood of each token in the input by masking one at a time, leading to much higher performance when compared to the use of the same model for fine-tuning on the target task.

In this work we combine the ideas of basing the assumptions on the results of experimental linguistics as well as employing a generative language model in an unconventional way, in order to benefit both from linguistic research and the data-driven approaches.

Causal Relation Classification Model

We give a brief description of the model architecture we employ for all experiment settings in the later sections. We treat COPA as a *sequence classification problem*, in which we want to predict for an input sentence s, consisting of a premise p and choice c, a label $l \in \{0, 1\}$, indicating whether p and c are connected by a *causal* relation. While the original COPA data contains both *cause* and *effect* relations, we make use of the same property observed by Li, Chen, and Van Durme (2019), namely that the relation between any premise p and continuation c is *reversible*. For example, for the premise-choice pair *"The woman's date wanted to look like a gentleman."*, *"He opened the door for her."* that is connected in an *effect* relation (*"The woman's date wanted to look like a gentleman,* ***therefore*** *he opened the door for her."*), we can treat the reverse relation as *causal*: *"He opened the door for her,* ***because*** *the woman's date wanted to look like a gentleman."*[1], by switching the order of premise and choice sentences. This simplifies our classification problem, as the model needs only learn to predict a single relation.

For our classification model, we use the implementation of RoBERTa (Liu et al. 2019) provided by Huggingface[2] to encode a given input sentence, with a simple linear classifier that takes the final encoding of RoBERTa's `<s>` token as input. Figure 1 shows a depiction of our model architecture. For all our experiments, we use *RoBERTa_large* as the encoder model, and a single-layer linear classifier with 1024-dimensional input and 2 output dimensions, with subsequent softmax.

During model training, we use the cross-entropy loss as optimization objective to predict $P(label|conv(p, c, r))$, where p and c are the original premise and a single choice, r is the original relation between p and c, $label \in \{0, 1\}$, and $conv(p, c, r)$ is a function that converts the premise and choice into an input suitable for RoBERTa, similar to Li, Chen, and Van Durme (2019):

$$conv(p, c, r) = \begin{cases} \text{<s> } w_1^p \ldots w_n^p \text{ because } w_1^c \ldots w_n^c \text{ </s>} \\ \quad\text{, if r = cause} \\ \text{<s> } w_1^c \ldots w_n^c \text{ because } w_1^p \ldots w_n^p \text{ </s>} \\ \quad\text{, if r = effect} \end{cases}$$

At inference time, we take as correct choice the one

[1] Arguably, this sentence is grammatically marked, as the pronoun appears before its antecedent, however we consider this drawback minor in comparison to the gain of unifying all examples into the same relation type.

[2] https://huggingface.co/

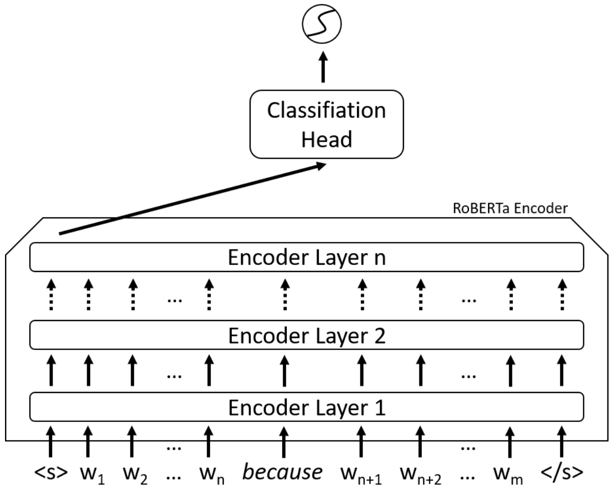

Figure 1: Our classification architecture used to predict COPA relations. A converted premise-choice sentence is encoded with a pre-trained RoBERTa model, and classified using a simple feed-forward network.

which, under the trained model, satisfies:

$$c^* = \underset{c \in \{\text{choice1, choice2}\}}{\arg\max} \quad P(1|conv(p, c, r))$$

Similarly to Kavumba et al. (2019), we use the learning rates of $1e-6$, $2e-6$ and $3e-6$, 20 different seeds per learning rate, weight decay of 0.01, and a batch size of 32. We train for a maximum of 50 epochs, stopping early when performance on the development set ceases to improve. We average the model performance on the development data over the results of the 20 seeds per learning rate, remove the bottom and top two outliers, and evaluate those models trained with the best-performing learning rate on the test set.

Wherever we train models on original or augmented datasets, we assume this very model architecture and training scheme. In addition, in result tables and discussions, we will refer to the original COPA training data as *Base*, for brevity.

Augmentation with Adversarial Examples

Given that the COPA dataset is very small, there is a need for methods that make models trained on it perform better with regard to both highest possible accuracy and robustness. This section presents work on using adversarial training by *perturbing* the original training set input sentences slightly, and using the sentences that make the trained model fail as additional data points.

Following the framework of Zang et al. (2020) we adapt their approach to the task of COPA with various modifications. Zang et al. (2020) work on the natural language inference and sentiment analysis tasks using a BERT-based model, adversarially attacking the inputs, and show that using the perturbed inputs as additional training data makes the BERT model more robust. The input perturbation consists of substituting some content words in the input with other words that share the same basic semantic units with the original word. The authors filter the potential substitution words by only using those that can appear as the same part of speech as the original word using HowNet (Qi et al. 2019) as a resource of word substitutions. Since computing all the combinations is intractable, Zang et al. (2020) then use a Particle Swarm Optimization (Kennedy and Eberhart 1995, PSO) algorithm to find the cases in which the model correctly classifies the original instance, but fails on the perturbed one.

In adapting this approach to the task at hand, we perform the following adjustments in our implementation:
Knowledge Base: we replace HowNet with WordNet (Fellbaum 2012) due to the resource size and quality.
Adversarial Attacks: we replace Particle Swarm with Ant Colony Optimization (Dorigo 1992, ACO) for choosing the best perturbation, with the aim of finding the most optimal adversarial examples. ACO is an algorithm inspired by biology – the way ants find their way around their environment – and belongs to the category of population-based search algorithms, which also includes PSO. It is a type of a meta-heuristics technique (Talbi 2009) with a focus on exploration space rather than optimization. ACO is able to find the *best path* between two points in a graph, which in our case are the first and final words in a sentence, and the paths are all the potential lexicalizations of each token;
Sentence Length: shorter sentences are included in the perturbations due to the short average length of COPA sentences (see also Table 2);
Semantic Substitution: potential substitution words are only selected if the substituting word has the same sense as the original one, in order to ensure that the words are only substituted with their synonyms which are relevant given the context, so as not to change the meaning of the sentence. To this end we perform Word Sense Disambiguation using SupWSD[3] introduced by Papandrea, Raganato, and Bovi (2017).
Pretrained Model: We replace the BERT model used by Zang et al. (2020) with RoBERTa as the base model, as described previously, due to it achieving higher results than BERT on a variety of Natural Language Processing tasks.

With these modifications, applying our adversarial attack method to the COPA training set leads to an 11.82% success rate in failing a previously trained RoBERTa-based COPA classification model where it originally classifies the inputs correctly. These successful attacks result in a set of 76 additional training items which we merge into the original training data, including changes in either the premise or one of the alternatives such as the following:

Original: The man craved a *cigarette*. He was addicted to nicotine.
Perturbed: The man craved a *butt*. He was addicted to nicotine.

While most generated perturbations seem to be of a high quality, there are occasional examples of either word sense

[3] https://supwsd.net/supwsd/

	#Train	Min	Max	Mean	Std
COPA Dev					
Base	400	82.80	88.80	86.21	1.84
+Adversarial	476	**85.80**	**89.20**	**87.35***	**1.10**
COPA Test					
Base	400	79.00	91.00	85.69	4.11
+Adversarial	476	**85.00**	**92.00**	**88.19***	**2.40**
Balanced COPA Dev					
Base	400	76.50	86.00	81.72	2.64
+Adversarial	476	**79.50**	86.00	**82.84***	**1.49**

Table 1: The results of the adversarially trained model compared to the baseline. * indicates a significant improvement over the Base model ($p < 0.001$).

disambiguation or subsequent replacement failing, for example in:

Original: The dog emitted a foul smell. The *skunk* sprayed the dog.
Perturbed: The dog emitted a foul smell. The *lowlife* sprayed the dog.

We leave the task of improving the automatic disambiguation to future work, but note that such blatantly wrong perturbations are a small minority of cases.

Adversarial Attack Evaluation

To evaluate the impact that our adversarially generated data has on model performance, we train classification models as described above on the base data set as well as on the data augmented with adversarial examples. Table 1 summarizes the results of model performance.

The adversarially enhanced models outperform the models trained on the original data alone in terms of both average performance and standard deviation. Hence, adversarial training leads to both higher performance and higher robustness of the model with as little as 76 additional data points. The improvements are statistically significant ($p < 0.001$) according to adapted approximate randomization (Noreen 1989) provided by Roemmele, Bejan, and Gordon (2011).

Augmentation by Causal Sentence Extraction and Distractor Generation

While the COPA dataset is very well curated, it only contains $1,000$ examples in total, providing an extremely low-resource setting, and generating more such high-quality data by hand would be quite slow and laborious. In this section we propose a method for automatically finding new data for augmentation from large free-form text. This augmentation involves three important steps, described in detail in the remainder of this section.

	Min	Max	Mean	Median	Std
Premise	2	13	6.2	6	1.8
Choice 1	2	11	5.1	5	1.6
Choice 2	2	11	5.1	5	1.6
Total Length	5	22	11.2	11	2.4
Premise/Total	22.2	82.8	54.7	54.5	0.1

Table 2: COPA dataset word counts. “Total Length” is averaged over all premise-choice pairs. The last row shows the ratios of premise to total length.

Data Augmentation Approach

Linguistically-motivated Filtering Strategy. As mentioned above, the original COPA dataset contains examples with *cause* and *effect* relations. In order to find relevant new sentences in large data, we require a filtering strategy that effectively extracts such *cause-effect* relations. To this end, we draw from linguistic theories to motivate strict criteria for filtering.

As with many discourse relations, *cause* and *effect* can be expressed either *implicitly* - through the content of the sentences, or *explicitly* - through the use of discourse connectives. In this work we rely on the explicit expression of such relations for finding causally linked clauses in raw web text. We define a total of 8 causal connectives[4], including forward- and backward-projecting instances.

While the defined connectives will yield generally appropriate sentences, the original COPA dataset imposes a number of additional restrictions on our desired augmented data. In particular, COPA contains premise-choice pairs that on average form relatively short sentences, and are balanced with regard to the lengths of the premise and the choices. Some length statistics of the original dataset are shown in Table 2. In accordance with these properties, we refine our filtering strategy to only consider sentences between 5 and 22 words long, with a connective in the center ±2 words. Finally, we observe that neither premises nor choices in COPA contain a (potential) *cause-effect* connective themselves. We therefore adapt our filter to reject sentences that contain more than one of our defined discourse connectives.

Furthermore, we also filter out vaguely phrased sentences such as “*My store doesn’t do that because that method can be unreliable.*”. The latter sentence does not contain much information, as it refers to actions and entities outside of the sentence, and only vaguely refers to the main events of the two clauses with verbs like “*do*” and “*be*”. In order to filter such sentences out, we only consider sentences which contain “*implicitly causal verbs*”. Implicit causality refers to a characteristic that certain verbs exhibit, namely that part of their meaning implies a causal directionality (Garvey and Caramazza 1974). For instance, the verb “*apologize*” is implicitly causal as it pertains to preferences to attribute the cause of the apology to the subject of the verb, meaning that when processing a sentence containing this verb, people tend to assume that the cause of the apology is an action per-

[4] as a result, because, if, since, so, therefore, thus, when

formed by the person apologizing. In this work, sentences are filtered out if they do not contain any implicitly causal verb as experimentally determined by Ferstl, Garnham, and Manouilidou (2011).

Large Free-form Text Resources. Since our criteria for candidate augmentation sentences are quite strict, we expect only a small amount of any available text resource to meet them. Additionally, sentences in COPA are quite varied in terms of their (perceived) source, ranging from sentences likely encountered in childrens' books to ones most likely sourced from newspaper text. Using traditional standard single-domain datasets for the augmentation task thus seems sub-optimal. We therefore opt to use the recently published OpenWebText corpus [5], itself derived from a non-open dataset introduced in Radford et al. (2019). OpenWebText contains ∼40GB of text from over 8 million documents, spanning a plethora of resources and domains. These properties make it an ideal resource for our strict COPA sentence filtering approach.

Data Augmentation Tool Chain. With a text resource and filtering strategy in place, we set up our tool chain for data augmentation as follows. First, we extract from OpenWebText a potentially large number of sentences that conform to the requirements set out before. To ensure the validity of the extracted sentences and of them encoding a *cause/effect* relationship, we then analyze each sentence using the Penn Discourse Treebank parser by Lin, Ng, and Kan (2014), and reject all sentences that either fail to parse, or do not parse as having the desired relation. After all filtering steps and PDTB parsing, we have obtained valid sentences constituting *positive* examples of a premise and a single choice.

To generate the distractor choices, we devise three strategies. For a given *premise-choice pair* we either (i) randomly choose the second choice from OpenWebText; (ii) choose it from OpenWebText if it shares at least one content word with the premise; (iii) use GPT-2 (Radford et al. 2019) to automatically generate a distractor. The (i) *random* selection of the distractor assumes a low probability of accidentally choosing a sentence that happens to be causally linked to our premise. While being the simplest heuristic, this method leads to examples which are potentially relatively easy to solve, as features such as sentence similarity could be expected to discriminate between the true cause and a random sentence from the web. The (ii) *overlap* selection method assumes that alternatives which overlap in their content words with the premise would be semantically more similar and therefore potentially more difficult for models to solve. Finally, the (iii) *GPT-2* method assumes the distinct denotations of certain discourse connectives, as discussed below.

For generating alternatives automatically with GPT-2, we make use of the same cause-effect reversibility discussed previously, re-writing all filtered augmentation data to encode backward causal relations. Then, the first part of the resulting sentence is selected up to but not including "*because*" and the conjunction connective is added as a new sentence start "*. And*" to condition the language model. The second half of the sentence is subsequently generated by GPT-2, subject to the aforementioned length constraints. We finally take the resulting continuation as the distractor choice of the augmented data point. The conjunction connective is chosen here because whenever it denotes temporal relations, they tend to be forward relations. That is, sentences of the form "*A and B*", where *A* and *B* are clauses, usually refer to chronologically ordered events, where *A* happens before *B*. In such cases, *B* can be an effect of *A*, however it is unlikely to be the cause of *A*. Such a forward relation can be employed here due to the fact that all the causal relations are converted into backward ones, which leads to the true and distractor alternatives denoting vastly different types of discourse relations - by type (causal vs. expansion) as well as direction (backward vs. forward). Thus, with this linguistically informed technique, we generate distractors which are semantically related to the premise based on the fact that they are generated as conjuncts to the premise. Between our proposed generative augmentation approaches, we hypothesize this to be the most difficult type of example for a causal inference model to distinguish, and thus conjecture it to yield the best results when used for data augmentation, compared to the other two methods.

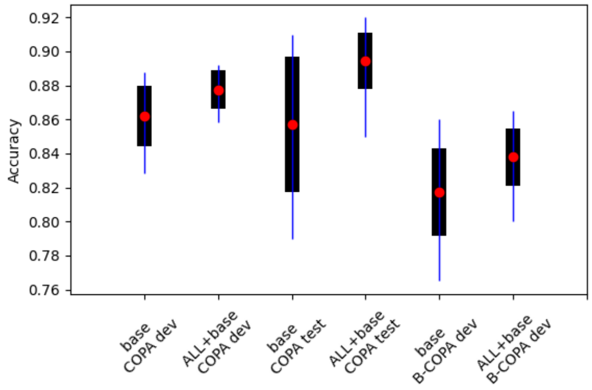

Figure 2: Results of the baseline model (trained on the original COPA training set) and the best performing model (trained on GPT-2 generated distractors together with the random and overlap methods in addition to the COPA training data), evaluated on the original COPA dev set, test set and the Balanced COPA dev set.

Experiments with Augmented Data

To evaluate the impact of our extractive augmentation approach on COPA performance, we carry out comparative experiments between setups using the different methods of generating new data.

We train our general classification architecture on data derived from the augmentation strategies described above. For this, we first generate 400 new training instances per generation method (random, overlap, and GPT-2), and train models on the original COPA data, individual augmented datasets, and the merged sets of augmented and original data. Table 3 shows results when training on the individual datasets, as well as on combined original and augmented data.

[5] http://Skylion007.github.io/OpenWebTextCorpus

	#Train	Min	Max	Mean	Std
COPA Dev					
Base	400	82.80	88.80	86.21	1.84
Random	400	60.40	72.40	68.51	3.18
+Base	800	84.80	88.80	86.73	1.41
Overlap	400	60.40	70.40	66.01	3.49
+Base	800	85.40	88.60	86.74	**1.07**
GPT-2	400	64.60	75.00	71.25	2.84
+Base	800	83.60	**89.20**	86.93*	1.75
All	1200	74.44	79.80	76.85	1.71
+Base	1600	**85.80**	**89.20**	**87.75***	1.16
COPA Test					
Base	400	79.00	91.00	85.69	4.11
Random	400	57.00	71.00	63.63	4.47
+Base	800	78.00	93.00	86.33	3.94
Overlap	400	57.00	73.00	65.19	4.65
+Base	800	**86.00**	90.00	88.13	**1.67**
GPT-2	400	62.00	77.00	71.31	3.65
+Base	800	**86.00**	**94.00**	**90.24***	2.28
All	1200	70.00	83.00	77.13	3.65
+Base	1600	85.00	92.00	89.44*	1.71
Balanced COPA Dev					
Base	400	76.50	86.00	81.72	2.64
Random	400	60.00	71.50	63.31	3.31
+Base	800	75.50	**86.50**	82.80	2.78
Overlap	400	57.00	71.50	64.16	3.51
+Base	800	**80.00**	85.00	82.03	**1.50**
GPT-2	400	66.00	72.50	69.31	2.15
+Base	800	79.50	86.00	**83.94***	1.91
All	1200	68.50	78.50	73.44	2.66
+Base	1600	**80.00**	**86.50**	83.78*	1.74

Table 3: The results of the model trained on data augmented through causal sentence extraction and distractor generation, compared to the baseline of trained only on original COPA data. * indicates a significant improvement over the Base model ($p < 0.001$).

Across the board, we observe that training with augmented data in addition to the original COPA data is able to boost model performance over training on just the Base data alone. It is also worth noting that all augmentation methods *on their own* seem to be able to provide at least a decent learning signal, with GPT-2 and all combined data even reaching scores up to and above the 70s on the dev, test, and balanced sets. When training with all augmentation data, models on average reach as high as 77% on the dev and test sets, and around 73% on balanced dev.

When adding the original COPA to the augmented data sets, we consistently achieve higher model performance. A single one of our models trained on COPA alone achieves 91% on the test set, however, this comes at the price of a severe standard deviation of over 4 points – a deviation only "topped" by models trained on the random or overlap based data *alone*. The single best model according to the test set is based on GPT-2-generated data in addition to the COPA training data, reaching 94% accuracy. When training on all data combined, the mean results on the COPA test set are only slightly lower. On the balanced COPA dev set, again, *GPT-2+Base* achieves the highest mean model performance, while the highest maximum is shared between the model based on all available data and the Random+Base model, while the highest minimum score is achieved by the model trained on all data. This goes to show that seemingly, our augmentation strategies successfully help overcome the base models' reliance on – and exploitation of – the superficial cues described previously, boosting performance on data that features no such cues.

One important improvement, also reflected in Figure 2, is that training with augmented data in addition to COPA greatly stabilizes model training, and consistently improves the models' standard deviation in almost all cases, especially on the test and balanced dev sets. Training on just the original COPA data in fact yields very unstable performance on those data sets, with a deviation of over 4% with individual evaluations ranging from 79% – 91% on test, and 76.5% – 86.00% on balanced dev. This renders actual model performance rather luck-based, relying heavily on randomness. With the use of augmented data, we are able to greatly stabilize the models. This leads to just over $1/3$ of COPA's original standard deviation on the test set, and just under one point decrease on the balanced dev set, when training with all available data. On the balanced set, our overlap-based augmentation almost halves the original standard deviation.

While all the augmented datasets (*Random+Base*, *Overlap+Base* and *GPT-2+Base*) lead to higher model performance and lower standard deviation scores in most settings, the *GPT-2+Base* augmentation outperforms the other two methods, and is the dataset that consistently yields significant improvements, on par with the models trained on all available data.

Overall, our experiments suggest that our augmentation strategies are able to generate data that contains high-quality "COPA-like" examples, useful for improving model performance over training with the hand-crafted and curated dataset alone.

Conclusions

In this work we have explored novel applications of linguistic knowledge to the task of detecting plausible causal relations between clauses. We have introduced a set of data augmentation techniques that can be applied by adversarially attacking existing models to find weak spots, or by using generative models to produce entirely new data. Our methods are able to generate high-quality data to augment the originally available training set of COPA. Experiments with data derived from our adversarial attack and various augmentation strategies show that our methods can help make model training more robust while also improving performance.

In the future, we would like to adapt our methods to further (low resource) data sets, to explore the general viability of our adversarial and augmentation strategies for improving model performance.

References

Asr, F. T.; and Demberg, V. 2013. On the Information conveyed by Discourse Markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, 84–93.

Blass, J. A.; and Forbus, K. D. 2017. Analogical Chaining with Natural Language Instruction for Commonsense Reasoning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota.

Dorigo, M. 1992. Optimization, Learning and Natural Algorithms. *PhD Thesis, Politecnico di Milano* .

Fellbaum, C. 2012. WordNet. *The encyclopedia of applied linguistics* .

Ferstl, E. C.; Garnham, A.; and Manouilidou, C. 2011. Implicit Causality Bias in English: A Corpus of 300 Verbs. *Behavior Research Methods* 43(1): 124–135.

Furbach, U.; Gordon, A. S.; and Schon, C. 2015. Tackling Benchmark Problems of Commonsense Reasoning. *Bridging@ CADE* 1412: 47–59.

Furbach, U.; and Schon, C. 2016. Commonsense Reasoning meets Theorem Proving. In *German Conference on Multiagent System Technologies*, 3–17. Springer.

Garvey, C.; and Caramazza, A. 1974. Implicit Causality in Verbs. *Linguistic inquiry* 5(3): 459–464.

Goodwin, T.; Rink, B.; Roberts, K.; and Harabagiu, S. 2012. UTDHLT: COPACETIC System for Choosing Plausible Alternatives. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 461–466. Montréal, Canada.

Goodwin, T. R.; and Demner-Fushman, D. 2019. Bridging the Knowledge Gap: Enhancing Question Answering with World and Domain Knowledge. *arXiv preprint arXiv:1910.07429* .

Gordon, A. S.; Bejan, C. A.; and Sagae, K. 2011. Commonsense Causal Reasoning using Millions of Personal Stories. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 1180–1185. San Francisco, California.

Iter, D.; Guu, K.; Lansing, L.; and Jurafsky, D. 2020. Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models. *arXiv preprint arXiv:2005.10389* .

Jabeen, S.; Gao, X.; and Andreae, P. 2014. Using Asymmetric Associations for Commonsense Causality Detection. In *13th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2014)*, 877–883. Gold Coast, QLD, Australia.

Kavumba, P.; Inoue, N.; Heinzerling, B.; Singh, K.; Reisert, P.; and Inui, K. 2019. When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 33–42. Hong Kong, China.

Kennedy, J.; and Eberhart, R. 1995. Particle Swarm Optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, 1942–1948. IEEE.

Li, Z.; Chen, T.; and Van Durme, B. 2019. Learning to Rank for Plausible Plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4818–4823. Florence, Italy.

Lin, Z.; Ng, H. T.; and Kan, M.-Y. 2014. A PDTB-styled End-to-End Discourse Parser. *Natural Language Engineering* 20(2): 151–184.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* .

Mtarji, A. 2019. *Commonsense reasoning using path analysis on semantic networks*. masterthesis, Universität Koblenz-Landau, Universitätsbibliothek.

Noreen, E. W. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Papandrea, S.; Raganato, A.; and Bovi, C. D. 2017. SupWSD: A Flexible Toolkit for Supervised Word Sense Disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 103–108. Copenhagen, Denmark.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A. K.; and Webber, B. L. 2008. The Penn Discourse TreeBank 2.0. In *LREC*. Citeseer.

Qi, F.; Yang, C.; Liu, Z.; Dong, Q.; Sun, M.; and Dong, Z. 2019. Openhownet: An Open Sememe-based Lexical Knowledge Base. *arXiv preprint arXiv:1901.09957* .

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683* .

Rahimtoroghi, E.; Hernandez, E.; and Walker, M. A. 2017. Learning Fine-grained Knowledge about Contingent Relations between Everyday Events. *arXiv preprint arXiv:1708.09450* .

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium 2011 - Logical Formalizations of Commonsense Reasoning*, 90–95. Stanford University, CA.

Shwartz, V.; West, P.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2020. Unsupervised Commonsense Question Answering with Self-Talk. *arXiv preprint arXiv:2004.05483* .

Siebert, S.; Schon, C.; and Stolzenburg, F. 2019. Commonsense Reasoning using Theorem Proving and Machine Learning. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 395–413. Canterbury, United Kingdom.

Talbi, E.-G. 2009. *Metaheuristics: From Design to Implementation*, volume 74. John Wiley & Sons.

Tamborrino, A.; Pellicano, N.; Pannier, B.; Voitot, P.; and Naudin, L. 2020. Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning. *arXiv preprint arXiv:2004.14074* .

Zang, Y.; Qi, F.; Yang, C.; Liu, Z.; Zhang, M.; Liu, Q.; and Sun, M. 2020. Word-level Textual Adversarial Attacking as Combinatorial Optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6066–6080. Seattle, Washington.