Guided Generation of Cause and Effect

Zhongyang Li[1,2], **Xiao Ding**[1*], **Ting Liu**[1], **J. Edward Hu**[2] and **Benjamin Van Durme**[2]
[1]Harbin Institute of Technology, China
[2]Johns Hopkins University, USA
{zyli,xding,tliu}@ir.hit.edu.cn, {edward.hu,vandurme}@jhu.edu

Abstract

We present a conditional text generation framework that posits sentential expressions of possible causes and effects. This framework depends on two novel resources we develop in the course of this work: a very large-scale collection of English sentences expressing causal patterns (**CausalBank**); and a refinement over previous work on constructing large lexical causal knowledge graphs (**Cause Effect Graph**). Further, we extend prior work in lexically-constrained decoding to support *disjunctive* positive constraints. Human assessment confirms that our approach gives high-quality and diverse outputs. Finally, we use CausalBank to perform continued training of an encoder supporting a recent state-of-the-art model for causal reasoning, leading to a 3-point improvement on the COPA challenge set, with no change in model architecture.

1 Introduction

Causal knowledge acquisition is crucial for various Artificial Intelligence tasks, such as causal event graph construction, reading comprehension and future event prediction. We propose an approach for acquiring causal knowledge through generating multiple plausible causes (reasons, explanations) and effects (results, consequences) for a provided input sentence. As exemplified in Figure 1, we develop two conditional decoders, one per causal direction. To train such models we mine a large-scale corpus of causal expressions from open domain web text, at a scale greatly surpassing prior work. Our goal is to generate multiple *distinct* possible causes and effects, where each generated sentence is not intended to be a paraphrase of other candidates. To support this output diversity when conditioned on a single shared input sentence, we turn to lexically-constrained decoding [Post and Vilar, 2018; Hu *et al.*, 2019a], which allows for efficiently forcing a model to produce output containing one or more provided phrases. Our constraints are derived from a resource we construct for this work, replicating a prior effort in lexicalized causal knowledge graph construction [Luo *et al.*, 2016]. This graph captures causal relations as a mapping across lexical types, lemma-to-lemma, but our goal is to generate naturalistic sentences with appropriately inflected morphology: we therefore develop an approach for *disjunctive* positive lexical constraints, where a decoder's output must contain one of a set of provided words or phrases. In our case, these are morphological variants of the same base lemma, but our approach should benefit other applications of lexically-constrained decoding.

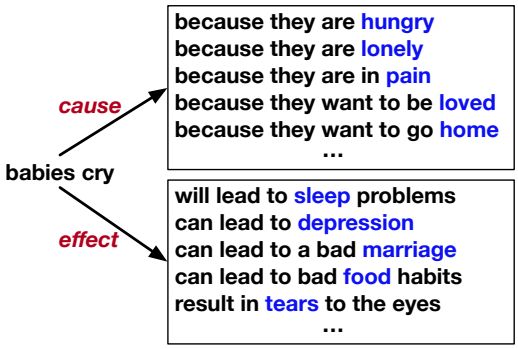

Figure 1: Possible causes and effects generated by our model, conditioned on the input sentence "babies cry". Tokens in blue are constraint keywords derived from our Cause Effect Graph, which are forced to be included in the outputs by constrained decoding.

While there is recent work in generating story endings conditioned on a context [Guan *et al.*, 2019; Wang and Wan, 2019; Luo *et al.*, 2019], such work does not require generated sentences to be strictly causes or effects. The ability to propose *explanations* for an input sentence by generating multiple causes and effects complements this emerging line of research. To our knowledge, this is the first work to consider open-ended generation of *causal sentences* at a large scale.

We evaluate through carefully designed human evaluation by comparing outputs from various baselines and our proposed model, finding that our model's outputs are preferred. We further demonstrate the usefulness of our new resource by taking a recent state-of-the-art causal reasoning system and boosting its results on the COPA test set by 3 points, relying only on continued training of the model's encoder. Our models and resources are made publicly available.[1]

In this paper, we make the following contributions:

*Corresponding author. Performed while the first author was visiting Johns Hopkins University.

[1]http://nlp.jhu.edu/causalbank

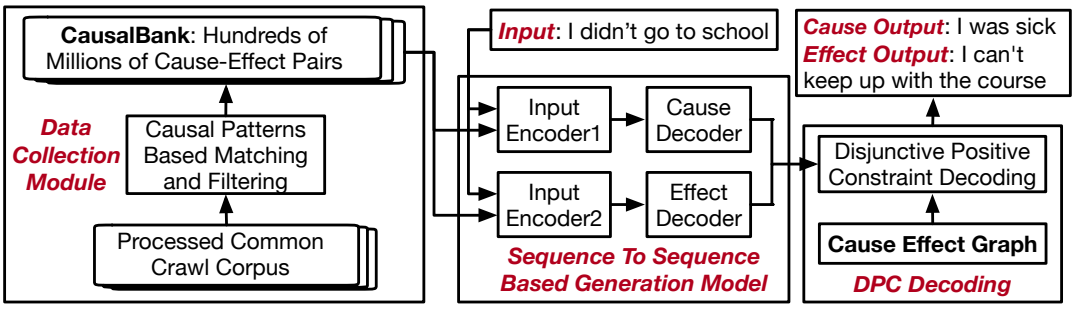

Figure 2: Our approach for generating plausible causes and effects.

Causal Pattern
as, as a consequence/result of, as long as, because, because of, caused by, due/owing to, in response to, on account of, result from
accordingly, consequently, bring on/about, give rise to, induce, in order to, lead to, result in, prevent/stop...from, and for this reason, cause, for the purpose of, if...then, ,_so, so that, thereby, therefore, thus, hence

Table 1: Causal patterns (their morphological variants are ignored) used to get the CausalBank corpus. The first row of patterns belong to the EPC category, while the second row belong to the CPE category.

- proposing the task of open causal generation: producing possible causes and effects for any free-form textual event;
- construction of a causal corpus (CausalBank) containing 314 million CE (cause-effect) pairs;
- an extension to lexically-constrained decoding that supports disjunctive positive constraints (DPC);
- human and automatic evaluations illustrating our method can generate high-quality and diverse causes and effects.

2 Approach

As shown in Figure 2, our proposed approach for open-ended causal generation includes a data collection module (Section 2.1), a Cause Effect Graph (Section 2.2), and two DPC (disjunctive positive constraint) decoding based Transformer encoder-decoder models (Section 2.3).

2.1 CausalBank: A Sentential Causal Corpus

Existing causal corpora were not built to support our goal for open-ended causal generation given any free-form textual input: as in neural machine translation (NMT), we need a large training set with millions of examples. Thus we harvest a large causal dataset from the preprocessed large-scale English Common Crawl corpus (5.14 TB) [Buck *et al.*, 2014]. The key guidelines of our dataset are as follows: 1) The causal relation is explicitly expressed in text with a causal pattern e.g. 'because'; 2) The 'cause' and 'effect' arguments must both appear in the same sentence; 3) The 'cause' and 'effect' arguments can be of any length of contiguous text without overlaps between them; 4) Negative causal relations are filtered.

We do not rely on a supervised text extractor to pick out specific sub-spans of a sentence that represent a cause-effect pairing between propositions.[2] We instead curate a series of patterns from previous studies [Mirza *et al.*, 2014; Luo *et al.*, 2016; Girju, 2003]. These patterns can be classified into two categories, according to how they are mostly used in language to convey a causal relation: 1. EPC (effect-pattern-cause) category: *I am very sad* BECAUSE **I lost my phone**; 2. CPE (cause-pattern-effect) category: **The earthquake** RESULTED IN *many deaths*. For EPC patterns, we simply take the text on the left of the pattern as effect, and take the text on the right of the pattern as cause. The case is reversed for CPE category patterns. These patterns (shown in Table 1) were applied to the Common Crawl corpus, followed by post-filtering: duplicate removal; filtering explicitly negated relations and verbs in passive voice; and restricting the cause and effect to each contain at least two tokens. This results in our **CausalBank** corpus, denoted here as $\mathcal{B}$, with 133 M EPC + 181 M CPE = 314 M (c, e) (c refers to cause and e refers to effect) pairs in total. We manually evaluated 1,000 randomly sampled sentences from the corpus and found that 95% conveyed a meaningful causal relation.

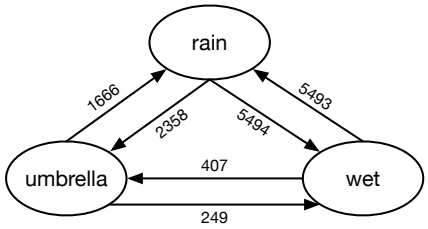

Figure 3: Cause Effect Graph: A lexical causal knowledge base.

2.2 Cause Effect Graph: A Lexical Causal KB

Following the method described in Luo *et al.* [2016] for creating a causal lexical knowledge base, we reproduce a variant of their **CausalNet** using the Common Crawl corpus [Buck *et al.*, 2014]. Given a sentence such as *"The storm **caused** a tremendous amount of damage on the landing beaches."*, this approach will harvest the lexical pairs (*storm*, *tremendous*), (*storm*, *amount*), (*storm*, *damage*), (*storm*, *landing*), and (*storm*, *beach*) as causal evidence. Stop words are removed and only pairs involving nouns, verbs, adjectives and

[2] We found poor annotator agreement on span boundaries in an initial investigation on crowdsourcing data for such a system; we intend to return to this in future work, investigating improvements to our results via trained extraction models for corpus pre-processing.

adverbs are retained. The extracted lexical pairs form a directed network of posited causal relations, where nodes in the network are lemmatized terms, and a directed edge between two terms indicates a causal relation, weighted by co-occurrence frequency. For comparison, Figure 3 gives a similar illustration as Figure 1 in Luo *et al.* [2016]. We refer to our artifact as a **Cause Effect Graph (CEG)**; Table 5 illustrates CEG contains more causal relations than CausalNet,[3] owing to the larger (5.14TB) and cleaner corpus used for extraction [Buck *et al.*, 2014].

2.3 Guided Generation

We use Sockeye [Hieber *et al.*, 2017] to train Transformer-based [Vaswani *et al.*, 2017] conditional generation models, one for causes, one for effects. Sockeye supports decoding via N-best (each step greedily chooses the top best N words in beam search based on the generated tokens) and random sampling (each step randomly sampling N words from the softmax distribution based on the generated tokens). The training data (CausalBank) is processed through Byte Pair Encoding [Sennrich *et al.*, 2016] to reduce vocabulary size.

Disjunctive Positive Constraints Decoding

Unlike in NMT, our intended outputs for a given input are diverse in meaning: we wish to generate multiple *semantically distinct* possible causes or effects. We induce diversity through hard lexical requirements during decoding, using causal keywords from our CEG as positive constraints on the output. A positive constraint forces the decoder to produce a sequence of tokens that contain the constrained sequence, which is achieved through a constrained beam search proposed by Post and Vilar [2018] and made efficient by Hu *et al.* [2019a].

Unfortunately, those prior works are restricted to *conjunctive* positive constraints: all items provided to the decoder *must* be present in the output. This is problematic in our case: our CEG maps lemmas to lemmas, and thus lemmas will form our constraints, but at generation time we do not require specific morphological inflections of our constrained terms. We wish not to constrain the decoder to a particular lemma, but to allow it to choose the best morphological form as appropriate in its context. For example, when generating a cause for *"I brought an umbrella"* with *rain* as the cause keyword, some valid cause sentences, e.g., *"It rained"* or *"It was a rainy day."*, would not be permitted based on prior work. One may circumvent this limitation by enumerating all morphological variants of a term, then apply each in turn as a positive constraint in distinct decoding passes. However, this approach does not scale, as its run-time grows exponentially in the number of initial constraints, each with multiple morphological variants.

Here we propose a solution of *disjunctive* positive constraint decoding, where each constraint is represented by a set of token sequences, and the decoder needs to include only one sequence from each set of constraints in the final output. We modify the algorithm from Hu *et al.* [2019a] to allow the decoder to explore the disjunctively constrained space in a single forward sequence, without significant computational overhead.

[3] 89.1M in contrast to 13.3M, with relations with a frequency of 5 or lower removed.

Algorithm 1 Decoding with Disjunctive Positive Constraints. We consider the generation of one sentence with a beam size of 1 for simplicity. Note that while a beam size of 1 reduces the constrained beam search, the handling of DPC is not affected.

input: a set of disjunctive constraint sets t, for each set s in t, $s_i = \{s_i^0, s_i^1, ..., s_i^n\}$ and $s_i^n = (w_i^{(n)(0)}, w_i^{(n)(1)}, ..., w_i^{(n)(m)})$ where $w_i^{(n)(m)}$ is the m^{th} token in s_i^n, one of the sequences of the disjunctive constraint set s_i
output: a token sequence $o = (o_0, o_1, ..., o_k)$
$trie := BuildTrie(\{s_0^0, ..., s_i^n\})$
while o_{k-1}! =EOS and $k < k_{max}$ **do**
  $o_k := ConstrainedBeamSearch((o_0, ..., o_{k-1}), t)$
  if o_k finishes the sequence s_p^q **then**
    for s_p^i in s_p **do**
      $trie := trie.prune(s_p^i)$
    end for
    Remove s_p from t
  end if
  $k := k + 1$
end while
return $(o_0, o_1, ..., o_k)$

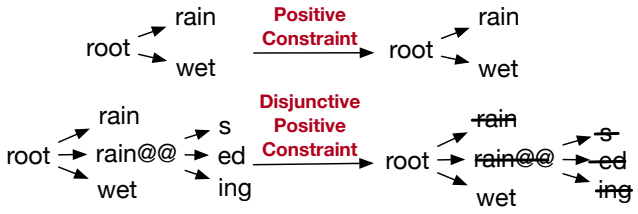

Figure 4: Trie states in positive constraint and disjunctive positive constraint, after generating the token '**rained**' in beam search.

In that work, constraints are represented in a trie, where each constraint is represented by a path from the root to a leaf. One or more state pointers are used to track how many tokens have been generated for each constraint, and tokens that induce more progress are prioritized in a modified beam search proposed by Post and Vilar [2018]. When a constraint is satisfied, the algorithm prunes the path representing that constraint. The distinguishing property of a disjunctive constraint is that once a sequence in a disjunctive set is satisfied, others in the set are also removed and no longer constraints.

For decoding with disjunctive constraints, we represent all constrained sequences, whether they are from the same disjunctive set or not, on a single trie. When a sequence is generated, we prune *all* sequences in the set as opposed to just the generated sequence. This modification gives us an efficient algorithm for applying disjunctive constraints, as illustrated in Algorithm 1 and Figure 4. While here we use morphological variants in our disjunctive set, our algorithm is broadly applicable for constraining on a set of synonyms or different subword segmentations of the same sequence.

Outputs Reranking While DPC decoding supports arbitrary number of disjunctive constraints in one beam search process, in practice only a few preferred constraints under the model will dominate any N-best output. To encourage diver-

Method	Dataset	Cause		Effect	
		Per	Acc	Per	Acc
RNN-LSTM	CB_10M	66.0	29.6	55.2	32.2
RNN-GRU	CB_10M	67.6	29.5	48.0	33.7
CNN	CB_10M	37.6	36.1	39.5	35.4
Conv-Transformer	CB_10M	29.5	38.9	31.1	38.2
Transformer	CB_10M	**28.3**	**39.1**	**29.9**	**38.4**
Transformer	CB_all	31.4	38.0	27.6	39.7
Transformer_BIG	CB_all	**29.9**	**38.5**	**26.4**	**39.8**

Table 2: Dev-set results: perplexity (**Per**), word accuracy (**Acc** (%)).

sity we first select a set of candidate constraint tokens from CEG, generate outputs per constraint, then merge and rerank the results. For example, if generating causes for the input sentence i = *"babies cry"*, we lemmatize each word in the sentence (*baby* and *cry*). These terms map to a set of lemmas via CEG, each associated with an observed frequency; we take the N-most frequent (highest weighted) such candidates: $t = \{w_1, w_2 ... w_N\}$. For each token w_i in t, such as *'love'*, we get a set of its morphological variants s_i ={*'love', 'loves', 'loved', 'loving'*} via the python package `patterns.en`, and pass s_i as a DPC, keeping the top M outputs. In total we derive $N * M$ (N=300 and M=5) sentences via N beam search decodings. These sentences are ranked by their associated negative log-likelihood scores, and we return the top K.

3 CausalBERT

Previous studies [Phang *et al.*, 2018; Li *et al.*, 2019b] have shown that applying intermediate auxiliary task training to an encoder such as BERT can improve performance on a target task. We designed an intermediate task for BERT using CausalBank $\mathcal{B}$, employing margin loss [Li *et al.*, 2019a; Li *et al.*, 2018a] in the objective function: $L(\Theta) = \sum_{(c,e)\in\mathcal{B}}(\max(0, m - f(c,e) + f(c',e'))) + \frac{\lambda}{2}||\Theta||^2$, where $f(c,e)$ is the score of true CE pair given by BERT model, $f(c',e')$ is the score of corrupted CE pair by replacing c or e with randomly sampled negative cause c' or effect e' from other examples in $\mathcal{B}$. $m > 0$ is the margin loss function parameter, which is set to 0.3. Θ is the set of BERT model parameters. λ is the parameter for L2 regularization, which is set to 0.00001.

By training BERT with this intermediate supervised task, we expect the model to acquire enhanced knowledge about the meaning of a causal relation, and can have better performance on downstream causal inference tasks.

4 Evaluation

We evaluate our proposed causal generation approach by both human and automatic metrics, and evaluate CausalBank by applying CausalBERT to COPA, which requires the model to choose the correct cause or effect from two candidates.

Model Selection

We first experiment on a small subset of our CausalBank corpus (CB_10M) – 10 million CE pairs from the causal pattern 'because' – considering different NMT encoder and decoder architectures (LSTM, CNN, Conv-Transformer [Gehring *et al.*, 2017], and Transformer).[4] For the cause generation model, e is used as the source and c is used as the target, which is reversed in training the effect model. Perplexity (Per) and word accuracy (Acc) are used to evaluate the model's performance. We find that Transformer constantly achieves the best performance (Table 2).

	Method	Cause				Effect			
		P@1	P@3	H	Div	P@1	P@3	H	Div
TrainSub	KNN	**89.0**	67.3	**0.85**	0.11	**98.0**	71.3	**0.90**	**0.02**
	GPT-2	31.0	22.3	0.39	0.13	8.0	9.3	0.30	0.11
	N-Best	59.0	45.3	0.53	0.15	63.0	42.7	0.53	0.11
	Random	68.0	59.3	0.66	0.11	74.0	61.7	0.70	0.09
	CN-Cons	72.0	71.3	0.79	**0.02**	66.0	67.0	0.76	**0.02**
	Gold-Cons	78.0	**75.3**	0.83	0.12	71.0	**73.0**	0.80	0.10
COPA_Dev	KNN	10.0	8.0	0.53	0.10	4.0	2.7	0.26	**0.01**
	GPT-2	40.0	34.0	0.45	0.12	38.0	32.0	0.46	0.10
	Random	66.0	53.7	0.65	0.09	62.0	46.7	0.57	0.08
	N-Best	69.0	65.0	0.77	0.08	**72.0**	68.0	0.82	0.07
	CN-Cons	**74.0**	70.0	0.81	**0.02**	**72.0**	**72.0**	**0.87**	0.02
	Gold-Cons	73.0	**73.0**	**0.87**	0.09	**72.0**	71.3	**0.87**	0.09

Table 3: Human evaluation results of cause and effect generation.

Then we train two versions of Transformer on the whole CausalBank corpus (CB_all). The small model's encoder and decoder both have 6 layers, with a hidden size and embedding size of 512. The big model's encoder and decoder have 12 layers and 4 layers, with a hidden size and embedding size of 768, leading to 134M parameters in total. The vocabulary size is 15,000. The training is stopped when the validation loss stagnates for 20,000 batches. For the cause generation model, e and c from only the EPC category (c, e) pairs are used as the source and target. For the effect generation model, c and e from only the CPE category (c, e) pair is used as the source and target. This setting always generates the right part of the sentence conditioned on the left part, which we find to give more reasonable outputs than the above architecture exploration experiments. The bottom of Table 2 shows the large Transformer model constantly achieves the best performance on development set, which contains 5,000 CE pairs.

Evaluating Generation

We evaluate the large Transformer model via human assessment, on two kinds of test sets. The first kind of test sets (TrainSub) contains 100 randomly sampled input examples from the model's training data. The second kind of test sets (COPA_Dev) contains 100 randomly sampled examples from the development set of COPA [Roemmele *et al.*, 2011] dataset, which are manually created gold sentences and never seen during the model's training stage.

The compared methods include a simplified KNN method (when the input is "babies cry", we match sentences exactly containing the input as the retrieved neighbors, e.g. "those babies cry loudly", and get the corresponding causes and effects), the GPT-2 124M language model [Radford *et al.*,

[4]Each of these models' encoder and decoder use the same architecture, e.g. both are 6-layer LSTMs, with a hidden size and embedding size of 512. All models are trained for 10 epochs. The vocabulary size is 10,000.

Method	Acc (%)
PMI [Jabeen *et al.*, 2014]	58.8
PMI_EX [Gordon *et al.*, 2011]	65.4
CS [Luo *et al.*, 2016]	70.2
CS_MWP [Sasaki *et al.*, 2017]	71.2
Google T5-base [Raffel *et al.*, 2019]	71.2
BERT-base [Li *et al.*, 2019a]	75.4
CausalBERT-base (ours)	**78.6**
Google T5-11B [Raffel *et al.*, 2019]	94.8

Table 4: Results on COPA-Test, contrasting prior results to a model by Li *et al.* built atop BERT-base. This model is improved by 3 points through adoption of CausalBERT.

2019] which can generate continuations conditioned on a start sentence (e.g. "babies cry because"), random sampling based decoding, N-best decoding, DPC decoding with constraint tokens from CEG (CN-cons), and DPC decoding with gold answer as constraint tokens (Gold-cons).

Four graduate students from the NLP field were used in annotation. Each was asked to give a score from $\{0, 1, 2\}$ for the generated {input, cause/effect} pair, where the guidelines are (take cause generation for example): if the generated answer does not make sense or can never be a reasonable cause, reason or explanation for the input event, give a score of 0; if the generated answer has grammatical errors but can be a reasonable cause, reason or explanation for the input event under some rare conditions (or beyond commonsense), give a score of 1; if the generated answer is a fluent sentence and can be a reasonable cause, reason or explanation with high probability, give a score of 2. Each pair was labeled by two annotators, and we average the judgments over two annotators per pair. The cohen's kappa score is 0.53.

Table 3 shows the human evaluation results. Three metrics are adopted: Precision at 1 **P@1** (an average score of 1.5 or above is seen as a valid causal answer); **P@3**; and average human score for each evaluated pair (**H**). For the TrainSub test set, the KNN method shows strong performance, especially for P@1 and the human scores. However, KNN performs worse for P@3, due to the absence of many possible answers for the same input. Meanwhile, our two versions of DPC decoding strategies (CN-cons, Gold-Cons) also show relatively better performance compared to other generation methods (GPT-2, Random and N-best decoding). KNN performs poorly on the COPA dev set, because most of the inputs never appear in the training data. However, CN-Cons and Gold-Cons can still achieve good performance.

Lexical Diversity We used a modified BLEU score to evaluate lexical diversity (**Div** in Table 3) where a lower score means a greater lexical diversity. Specifically, we calculate the associated BLEU-1 score between the gold answers and the generated top 3 outputs without brevity penalty. This modification ensures that we don't reward shorter outputs. In most cases, CN-Cons gets the lowest **Div** scores, showing that our DPC decoding and constraint tokens from CEG together, allows us to explore more in the causes and effects space, and generate more diverse outputs. Also we find that all of these BLEU scores are very low, compared with the BLEU scores in previous text generation studies [Hu *et al.*, 2019b; Vaswani *et al.*, 2017]. This is because our generation task is open-ended (as illustrated in Figure 1).

Evaluating CausalBank Table 4 shows our CausalBERT results on COPA test. Compared with prior strong knowledge-driven baseline methods, a BERT-base model trained with a margin-based loss [Li *et al.*, 2019a] achieved good performance. Following the experimental settings of Li *et al.* [2019a], when training the BERT-base model with additional CE pairs from CausalBank, we get an improvement of 3.2%, from 75.4% to 78.6%, showing that our corpus successfully augments BERT base to make it better for causal inference, which is a sign the corpus contains useful causal knowledge. We find that the number of CE pairs in the intermediate task matters: performance first improves and then decreases, with more training data added. [5] We get the best performance of 78.6% with 40 K training CE pairs. Though our result still has a big gap from the current SOTA performance on COPA (94.8% from the largest google T5-11B model), the intent of our experiment is just to illustrate how the only difference was in altering the pre-training with CausalBank. One could possibly get a SOTA model based on our corpus and the google T5 model, if publicly available.

5 Related Work

Conditional Text Generation Such efforts cover a large body of work, including machine translation, response generation and paraphrase generation. Most related is conditional story generation [Guan *et al.*, 2019; Wang and Wan, 2019; Luo *et al.*, 2019; Li *et al.*, 2018b], which aims to generate story continuations based on a given context. These works do not require generated sentences to be strictly causes or effects.

For causal generation, Rashkin *et al.* [2018] aimed to generate the likely intents and reactions of the event's participants, given a short free-form textual event. Sap *et al.* [2019] trained a multi-task model for fine-grained kinds of *If-Then* commonsense reasoning. However, the causal semantics considered in their work are restricted to a narrow space, and their models are trained on no more than one million examples. Further, their resource was based-on crowdsourcing, which carries risks of human bias [Rudinger *et al.*, 2017; Poliak *et al.*, 2018]. We harvest a significantly larger, open coverage causal corpus,[6] related in approach to DisSent [Nie *et al.*, 2019] but larger, focused on causality, and aimed primarily at generation rather than sentence representation learning.

Of various efforts in guided generation [Ammanabrolu *et al.*, 2019; Tang *et al.*, 2019; Clark *et al.*, 2018; Hu *et al.*, 2019b],

[5] This was not observed in related studies [Phang *et al.*, 2018; Li *et al.*, 2019b], where all training examples from the Multi-NLI dataset were used as an intermediate task. Similar behavior was observed in NMT in continued training for domain adaptation [Thompson *et al.*, 2019]. We believe ours to be a similar setting, where the "in-domain" causal data overwhelms the benefits of pretraining; adapting strategies from Thompson *et al.* is an avenue for future work.

[6] While we avoid pitfalls of elicitation, we acknowledge that like any corpus-extracted resource ours may suffer from *reporting bias* [Gordon and Van Durme, 2013]: some types of causes or effects that are known to humans but rarely or ever explicitly stated.

Sentential Causal Resource	**# CE Pairs**
TCR [Ning *et al.*, 2018]	172
SemEval-2007 Task4 [Girju *et al.*, 2007]	220
Causal-TimeBank [Mirza *et al.*, 2014]	318
CaTeRS [Mostafazadeh *et al.*, 2016]	488
EventCausalityData [Do *et al.*, 2011]	580
RED [O'Gorman *et al.*, 2016]	1,147
SemEval2010 Task8 [Hendrickx *et al.*, 2009]	1,331
BECauSE 2.0 [Dunietz *et al.*, 2017b]	1,803
EventStoryLine [Caselli and Vossen, 2017]	5,519
PDTB 2.0 [Prasad *et al.*, 2008]	8,042
Altlex [Hidey and McKeown, 2016]	9,190
PDTB 3.0 [Webber *et al.*, 2019]	13 K
DisSent [Nie *et al.*, 2019]	167 K
CausalBank (Ours)	**314 M**
Causal Knowledge Graph	**# CE Edges**
Event2mind [Rashkin *et al.*, 2018]	25 K
ConceptNet 5.7 [Speer *et al.*, 2017]	473 K
ASER Core [Zhang *et al.*, 2019]	494 K
Atomic [Sap *et al.*, 2019]	877 K
CausalNet [Luo *et al.*, 2016]	13.3 M
Cause Effect Graph (Ours)	**89.1 M**

Table 5: Contrasting size with example prior works: only the causal portion of these corpora are listed. The top are sentential causal corpora, while the bottom are graph-structure causal knowledge bases.

lexically-constrained decoding [Hokamp and Liu, 2017] is a modification of beam search originating in neural machine translation which allows the user to specify tokens that must (or must not) appear in the decoder's output.

Post and Vilar [2018] proposed a variant of lexically-constrained decoding that reduced complexity from linear to constant-time, which was made more efficient by Hu *et al.* [2019a]. We introduce an extension to lexically-constrained decoding that supports disjunctive positive constraints for multiple optional constraint keywords.

Sentential Causal Resources Existing causal corpora differ in their annotation guidelines and how they are constructed: (1) whether they consider only explicit or also implicit causal relations; (2) whether they consider only intra-sentence relations or if relations can cross sentences; (3) whether the annotation unit is word level or sentence level; and (4) whether the corpus is constructed automatically or by human effort. Ours is concerned with explicit only relations, within a single sentence, relating one part of a sentence to another, and employs constructed patterns but not sentence-level human annotation.

Already mentioned are recent crowdsourcing efforts [Rashkin *et al.*, 2018; Sap *et al.*, 2019]. More related are PDTB [Prasad *et al.*, 2008] and BECauSE [Dunietz *et al.*, 2017b], but where our resource goal is a much larger corpus, for the purpose of training a neural text generation model. Most related would be the extractive approach of DisSent [Nie *et al.*, 2019], but where we focus specifically on causality, and derive a much larger corpus. [Bethard and Martin, 2008] tagged a small corpus of event pairs conjoined with "and" as causal or not causal. CaTeRS [Mostafazadeh *et al.*, 2016] included causal relations from a commonsense reasoning standpoint. Richer Event Description [O'Gorman *et al.*, 2016] integrates real-world temporal and causal relations between events into a unified framework. Table 5 contrasts the size of causal portion of prior resources with our own.

Lexical Causal Resources Lexical semantic resources may encode causal properties on verbs (e.g., [Schuler, 2005; Bonial *et al.*, 2014]) and prepositions (e.g., [Schneider *et al.*, 2015]). Force dynamics theory [Talmy, 1988] from cognitive psychology posits three primary kinds of causal semantics [Wolff, 2007] – CAUSE, ENABLE and PREVENT – which were lexicalized as causal verbs [Wolff and Song, 2003]. The annotation scheme of Dunietz *et al.* [2017b] distinguishes three types of causal semantics: CONSEQUENCE, MOTIVATION, and PURPOSE. In PDTB 2.0 [Prasad *et al.*, 2008], "CONTINGENCY" has two subtypes ("Cause" and "Condition"). FrameNet [Baker, 2014] represents causal relations through a variety of unrelated frames (e.g., CAUSATION and THWARTING) and frame roles (e.g., PURPOSE and EXPLANATION). These efforts motivate our own causal patterns, categorized into: **CAUSE** (e.g. cause, result in, lead to), **EXPLANATION** (e.g. because, due to), **CONDITION** (e.g. if-then, as long as), **PURPOSE** (e.g. in order to, for the purpose of), and **PREVENTION** (e.g. stop/prevent-from).

Causal Knowledge Acquisition Causal knowledge acquisition [Radinsky *et al.*, 2012; Radinsky and Horvitz, 2013] is crucial for many AI systems, and it is often acquired via text. Hashimoto *et al.* [2014] and Kruengkrai *et al.* [2017] applied supervised learning techniques using a benchmark training data with over 100K human-annotated CE pairs. Dasgupta *et al.* [2018] explored general causal extraction using 5,000 labelled sentences. Do *et al.* [2011] is an example of a minimally supervised approach. Recent studies [Dunietz *et al.*, 2017a; Dunietz *et al.*, 2018] explored new supervised approaches on the BECauSE 2.0 [Dunietz *et al.*, 2017b] corpus.

Church and Hanks [1990] proposed the use of pointwise mutual information (PMI) for mining patterns via text co-occurrence. Many works have followed this strategy, e.g. [Chambers and Jurafsky, 2008; Riaz and Girju, 2010; Gordon *et al.*, 2011; Do *et al.*, 2011; Luo *et al.*, 2016]. Others have mined patterns via discourse patterns in the form of 'A led to B', 'if A then B', etc., e.g., [Khoo *et al.*, 2000; Girju, 2003; Zhao *et al.*, 2017]). See Asghar [2016] for review. Such efforts relate most closely to our CEGraph component, rather than our overall framework. Our concern is the generation of diverse potential causes and effects as natural language statemnts.

6 Conclusion

We investigate open causal generation for free-form textual input, and build a large sentential causal corpus which we used to train a generative model. We introduced a novel extension to lexically-constrained decoding that supports disjunctive positive constraints, where generated output is forced to contain one of a set of candidates. Automatic and human evaluations show that our method can generate high-quality and diverse causes and effects for new inputs.

Acknowledgements

We acknowledge the support of the National Key Research and Development Program of China (SQ2018AAA0101901),

the National Natural Science Foundation of China (NSFC) via Grant 61976073 and 61702137; the China Scholarship Council; and DARPA KAIROS (Hu and Van Durme).

References

[Ammanabrolu *et al.*, 2019] P. Ammanabrolu, E. Tien, W. Cheung, Z. Luo, W. Ma, L. Martin, and M. Riedl. Guided neural language generation for automated storytelling. In *Storytelling Workshop*, 2019.

[Asghar, 2016] N. Asghar. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*, 2016.

[Baker, 2014] C. Baker. Framenet: a knowledge base for natural language processing. In *Frame Semantics*, 2014.

[Bethard and Martin, 2008] Steven Bethard and James H Martin. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 177–180. Association for Computational Linguistics, 2008.

[Bonial *et al.*, 2014] C. Bonial, J. Bonn, K. Conger, J. D Hwang, and M. Palmer. Propbank: Semantics of new predicate types. In *LREC*, 2014.

[Buck *et al.*, 2014] C. Buck, K. Heafield, and B. van Ooyen. N-gram counts and language models from the common crawl. In *LREC*, 2014.

[Caselli and Vossen, 2017] T. Caselli and P. Vossen. The event storyline corpus: A new benchmark for causal and temporal relation extraction. 2017.

[Chambers and Jurafsky, 2008] N. Chambers and D. Jurafsky. Unsupervised learning of narrative event chains. In *ACL*, pages 789–797, 2008.

[Church and Hanks, 1990] K. Ward Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 1990.

[Clark *et al.*, 2018] E. Clark, Y. Ji, and N. A Smith. Neural text generation in stories using entity representations as context. In *NAACL*, pages 2250–2260, 2018.

[Dasgupta *et al.*, 2018] T. Dasgupta, R. Saha, L. Dey, and A. Naskar. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Annual SIGdial Meeting on Discourse and Dialogue*, 2018.

[Do *et al.*, 2011] Q. Do, Y. Chan, and D. Roth. Minimally supervised event causality identification. In *EMNLP*, pages 294–303, 2011.

[Dunietz *et al.*, 2017a] J. Dunietz, L. Levin, and J. Carbonell. Automatically tagging constructions of causation and their slot-fillers. *TACL*, pages 117–133, 2017.

[Dunietz *et al.*, 2017b] J. Dunietz, L. Levin, and J. Carbonell. The because corpus 2.0: Annotating causality and overlapping relations. 2017.

[Dunietz *et al.*, 2018] J. Dunietz, J. G Carbonell, and L. Levin. Deepcx: A transition-based approach for shallow semantic parsing with complex constructional triggers. In *EMNLP*, 2018.

[Gehring *et al.*, 2017] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. In *ICML*, 2017.

[Girju *et al.*, 2007] R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. T., and D. Y. Semeval-2007 task 04: Classification of semantic relations between nominals. 2007.

[Girju, 2003] Roxana Girju. Automatic detection of causal relations for question answering. 2003.

[Gordon and Van Durme, 2013] J. Gordon and B. Van Durme. Reporting bias and knowledge acquisition. In *AKBC*, 2013.

[Gordon *et al.*, 2011] A. S Gordon, C. A Bejan, and K. Sagae. Commonsense causal reasoning using millions of personal stories. In *AAAI*, 2011.

[Guan *et al.*, 2019] J. Guan, Y. Wang, and M. Huang. Story ending generation with incremental encoding and commonsense knowledge. In *AAAI*, 2019.

[Hashimoto *et al.*, 2014] C. Hashimoto, K. Torisawa, et al. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *ACL*, 2014.

[Hendrickx *et al.*, 2009] I. Hendrickx, Su Nam Kim, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *SE*, 2009.

[Hidey and McKeown, 2016] Christopher Hidey and Kathy McKeown. Identifying causal relations using parallel wikipedia articles. In *ACL*, pages 1424–1433, 2016.

[Hieber *et al.*, 2017] F. Hieber, T. Domhan, et al. Sockeye: A toolkit for neural machine translation. *arXiv*, 2017.

[Hokamp and Liu, 2017] Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In *ACL*, 2017.

[Hu *et al.*, 2019a] J E. Hu, Huda Khayrallah, et al. Improved lexically constrained decoding for translation and monolingual rewriting. In *NAACL*, 2019.

[Hu *et al.*, 2019b] J E. Hu, R. Rudinger, M. Post, and B. Van Durme. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *AAAI*, 2019.

[Jabeen *et al.*, 2014] S. Jabeen, X. Gao, and P. Andreae. Using asymmetric associations for commonsense causality detection. In *PRICAI*, 2014.

[Khoo *et al.*, 2000] Christopher SG Khoo, Syin Chan, and Yun Niu. Extracting causal knowledge from a medical database using graphical patterns. In *ACL*, 2000.

[Kruengkrai *et al.*, 2017] C. Kruengkrai, K. Torisawa, et al. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *AAAI*, 2017.

[Li *et al.*, 2018a] Z. Li, X. Ding, and T. Liu. Constructing narrative event evolutionary graph for script event prediction. In *IJCAI*, pages 4201–4207, 2018.

[Li *et al.*, 2018b] Z. Li, X. Ding, and T. Liu. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training. In *Coling*, 2018.

[Li *et al.*, 2019a] Z. Li, T. Chen, and B. Van Durme. Learning to rank for plausible plausibility. In *ACL*, 2019.

[Li *et al.*, 2019b] Z. Li, X. Ding, and T. Liu. Story ending prediction by transferable bert. In *IJCAI*, 2019.

[Luo *et al.*, 2016] Z. Luo, Y. Sha, K. Q Zhu, S. Hwang, and Z. Wang. Commonsense causal reasoning between short texts. In *Knowledge Representation and Reasoning*, 2016.

[Luo *et al.*, 2019] F. Luo, D. Dai, P. Yang, T. Liu, B. Chang, Z. Sui, and X. Sun. Learning to control the fine-grained sentiment for story ending generation. In *ACL*, 2019.

[Mirza *et al.*, 2014] P. Mirza, R. Sprugnoli, et al. Annotating causality in the tempeval-3 corpus. In *CAtoCL*, 2014.

[Mostafazadeh *et al.*, 2016] N. Mostafazadeh, A. Grealish, et al. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Events*, 2016.

[Nie *et al.*, 2019] A. Nie, E. Bennett, and N. Goodman. DisSent: Learning sentence representations from explicit discourse relations. In *ACL*, 2019.

[Ning *et al.*, 2018] Q. Ning, Z. Feng, et al. Joint reasoning for temporal and causal relations. In *ACL*, 2018.

[O'Gorman *et al.*, 2016] T. O'Gorman, K. Wright-Bettner, and M. Palmer. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *CNS*, 2016.

[Phang *et al.*, 2018] J. Phang, T. Févry, and S. R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv:1811.01088*, 2018.

[Poliak *et al.*, 2018] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. Hypothesis only baselines in natural language inference. In *LCS*, 2018.

[Post and Vilar, 2018] M. Post and D. Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *NAACL*, 2018.

[Prasad *et al.*, 2008] R. Prasad, N. Dinesh, et al. The penn discourse treebank 2.0. In *LREC*, 2008.

[Radford *et al.*, 2019] A. Radford, J. Wu, et al. Language models are unsupervised multitask learners. 2019.

[Radinsky and Horvitz, 2013] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *WSDM*, 2013.

[Radinsky *et al.*, 2012] K. Radinsky, S. Davidovich, and S. Markovitch. Learning causality for news events prediction. In *WWW*, 2012.

[Raffel *et al.*, 2019] C. Raffel, N. Shazeer, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683*, 2019.

[Rashkin *et al.*, 2018] H. Rashkin, M. Sap, E. Allaway, N. A. Smith, and Y. Choi. Event2Mind: Commonsense inference on events, intents, and reactions. In *ACL*, 2018.

[Riaz and Girju, 2010] M. Riaz and R. Girju. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Semantic Comp.*, 2010.

[Roemmele *et al.*, 2011] M. Roemmele, C. A. Bejan, and A. S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI*, 2011.

[Rudinger *et al.*, 2017] R. Rudinger, Chandler May, and B. Van Durme. Social bias in elicited natural language inferences. In *Ethics in NLP*, 2017.

[Sap *et al.*, 2019] M. Sap, R. Le Bras, et al. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, 2019.

[Sasaki *et al.*, 2017] S. Sasaki, S. Takase, et al. Handling multiword expressions in causality estimation. 2017.

[Schneider *et al.*, 2015] N. Schneider, V. Srikumar, J. D Hwang, and M. Palmer. A hierarchy with, of, and for preposition supersenses. In *Linguistic Annotation*, 2015.

[Schuler, 2005] Karin Kipper Schuler. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005.

[Sennrich *et al.*, 2016] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *ACL*, 2016.

[Speer *et al.*, 2017] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017.

[Talmy, 1988] L. Talmy. Force dynamics in language and cognition. *Cognitive science*, 1988.

[Tang *et al.*, 2019] J. Tang, T. Zhao, et al. Target-guided open-domain conversation. In *ACL*, 2019.

[Thompson *et al.*, 2019] B. Thompson, J. Gwinnup, et al. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *NAACL*, 2019.

[Vaswani *et al.*, 2017] A. Vaswani, N. Shazeer, et al. Attention is all you need. In *NIPS*. 2017.

[Wang and Wan, 2019] T. Wang and X. Wan. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In *IJCAI*, 2019.

[Webber *et al.*, 2019] B. Webber, R. Prasad, A. Lee, and A. Joshi. The pdtb 3.0 annotation manual, 2019.

[Wolff and Song, 2003] P. Wolff and G. Song. Models of causation and the semantics of causal verbs. *CP*, 2003.

[Wolff, 2007] P. Wolff. Representing causation. *Journal of experimental psychology: General*, 2007.

[Zhang *et al.*, 2019] H. Zhang, X. Liu, et al. Aser: A large-scale eventuality knowledge graph. *arXiv*, 2019.

[Zhao *et al.*, 2017] S. Zhao, Q. Wang, S. Massung, et al. Constructing and embedding abstract event causality networks from text snippets. In *WSDM*, 2017.