doi:10.3233/SSW220021

A Joint Model for Detecting Causal Sentences and Cause-Effect Relations from Text

Tirthankar DASGUPTA [a,1], Abir NASKAR [a] and Lipika DEY [a] Mohammad SHAKIR [a]

[a] *TCS Research, India*

Abstract. Text documents are rich repositories of causal knowledge. While journal publications typically contain analytical explanations of observations on the basis of scientific experiments conducted by researchers, analyst reports, News articles or even consumer generated text contain not only viewpoints of authors, but often contain causal explanations for those viewpoints. As interest in data science shifts towards understanding causality rather than mere correlations, there is also a surging interest in extracting causal constructs from text to provide augmented information for better decision making. Causality extraction from text is viewed as a relation extraction problem which requires identification of causal sentences as well as detection of cause and effect clauses separately. In this paper, we present a joint model for causal sentence classification and extraction of cause and effect clauses, using a sequence-labeling architecture cascaded with fine-tuned Bidirectional Encoder Representations from Transformers (BERT) language model. The cause and effect clauses are further processed to identify named entities and build a causal graph using domain constraints. We have done multiple experiments to assess the generalizability of the model. It is observed that when fine-tuned with sentences from a mixed corpus, and further trained to solve both the tasks correctly, the model learns the nuances of expressing causality independent of the domain. The proposed model has been evaluated against multiple state-of-the-art models proposed in literature and found to outperform them all.

Keywords. Causal sentence detection, Causal information extraction, Joint modeling, Evaluation

1. Introduction

Detecting causal information from text documents is an important language processing task that has a wide range of applications. Causal information abounds in scientific articles that report reasons behind various observed phenomena, along with details of the study based on which the conclusions are obtained. Similarly, analyst notes contain explanations about various economic or political phenomenon. Mining these rich repositories of documented knowledge provides valuable authoritative information that can be used for downstream decision making applications. Causality extraction from text is

[1] Corresponding Author.

rapidly gaining speed as the extracted cause-effect pairs are found to play a significant role in several downstream analytical and predictive tasks like identification of actionable items, question-answering, isolation of confounding variables and predictive variables for predictive systems [1,2]. Curating causal relations from text documents can also help in building repositories of causal insights which are useful for reasoning tasks [3].

The concept of causality can be informally introduced as a relationship between an antecedent e_1 and a consequence e_2, expressed as e_1 *causes* e_2 [4,5]. However, natural language texts contain an abundance of such relations appearing in different forms [6,7]. Even a single sentence expressing causal relations can be arbitrarily complex in structure, which makes the extraction task challenging. Below is a high-level categorization of causal sentences, depicted with examples which highlight the variability in the use of linguistic constructs.

- Single cause-single effect - Let us consider the sentence - *Leukocytosis is caused by bacterial infection.* In this sentence, Leukocytosis is cited as the effect while bacterial infection is cited as a cause. The positions of the cause and effect can be inter-changed as shown in the following example - ”Intravenous azithromycin causes ototoxicity”, where the causal phrase *Intravenous azithromycin* appears before the effect, *ototoxicity*.
- Multiple cause-multiple effect - The following sentence from a business news article - *The recent market falls have been the result of big budget deficits, as well as the US’s yawning current account gap*, depicts the presence of multiple causes for one effect.
- Causal chains - Sometimes a causal chain which expresses causal effects as a series of events can be observed, as illustrated by the following sentence: - *Unavailability of IT infrastructure support team has hugely impacted development and support activities thereby affecting customer relations.*

Causality is expressed within text documents in arbitrarily complex ways. The expression of causality may be implicit or explicit. Sometimes causal expressions are easy to detect as sentences that contain clauses related by “caused” or “because of”. However, there are many other ways in which causalities may be expressed, as illustrated in examples like “climate change induced by rise in greenhouse gases”, or “ribavirin was associated with lower incidence of acute respiratory distress syndrome”. Causality detection from text requires identification of causal sentences as well as isolation of the cause and effect phrases from it. It is often viewed as a relation classification and argument extraction problem. Causality detection from text is challenging since the number of sentences containing causal relations are few and far-between [8]. The number of non-causal sentences far outnumbers the causal sentences.

In this paper, we have proposed a joint model for causal sentence classification and cause-effect relation extraction using a multi-tasking architecture, which learns to do the tasks simultaneously. Most of the earlier works have focused on extracting the cause-effect components from a given causal sentence only. Our observation is that a joint modeling for classifying causal sentences and detection of the components can yield a more effective model, that also has more practical use. Another significant aspect of the proposed model is that, unlike many of its predecessors, the model doesn’t require the entities to be provided as input. Rather, the model learns to effectively isolate cause and effect *phrases*, and not just single words or entities. This is an important contribution of the present work, which is illustrated through the following example. For the sentence "The AIDS pandemic is caused by the spread of HIV infection", while most of the earlier models reported in literature finds "pandemic" as the cause and "infection" as effect, the proposed model can correctly identify "AIDS pandemic" as the cause and "HIV infection" as effect, which are obviously more precise. The third contribution of the paper lies in experimenting with cross-domain learning capability of the model. We show that by training the model over a mixed corpus, the model learns the essence of causality in a domain-independent way.

The joint model for causality detection and causal component identification is designed as a sequence-labeling architecture cascaded with fine-tuned Bidirectional Encoder Representations from Transformers (BERT) language model [9]. While a sentence classifier is trained to identify a causal sentence, the sequence labeling layer assigns a label cause (C), effect (E), causal connector (CC) or none (N) to each word. The labeling of connectives is a unique proposition of the work, which along with its companion cause and effect pair, helps in detection of causal relations from complex sentences more effectively. Most of the earlier work only identify causes and effects. As we will discuss later, the causal connectives play an important role in designing downstream applications with the extracted relations. The proposed model is also capable of extracting multiple causal relations from a single sentence in a more robust fashion by identifying all causal relations that bind cause-effect pairs through the causal connectives. The BERT-base model is fine-tuned with sentences from a mixed corpus, and further trained for learning both the above objectives. The fine-tuned model learns the nuances of scientific and/or business language, the typical events, entities and their states. This helps the subsequent layers learn the boundaries of causal phrases more effectively. The proposed model has been evaluated with multiple state-of-the-art baseline neural network architectures. Results show that it outperforms all the baseline models in most of the tasks.

The rest of the paper is organized as follows. Section 2 presents related work in this area. Section 3 presents the proposed joint model for causal sentence and cause-effect relation extraction. Section 4 provides details about datasets. Section 5 presents details of experiments and evaluations. Section 6 presents a short glimpse of how causal insights can be curated for insight generation. Finally section 7 concludes the paper.

2. Related Works

Early references to causality detection from text can be traced back to [10,11,12,13], who championed syntactic rule-based methods for detecting cause - effect pairs. ROTEUS [14] and COATIS [15] were two such systems designed using non-statistical techniques. The diversity of the causal expressions presented earlier clearly show that a syntactic rule based approach cannot detected all kinds of causal relations. Neural representations that can effectively learn both the semantic and syntactic characteristics of causality are expected to do a better job.

An early machine-learning based approach was proposed in [16], which defined the task as one of sequence labeling. The intent was to assign cause or effect labels to words in a sentence, depending on their roles. Khoo et al. in a series of works [17,10,18] reported extraction of explicit causal relations containing known causal connectives like "if-then" constructs, causative adverbs and adjectives etc. The method yielded very low precision of 19% on a test set of Wall Street Journal articles.

Increased focus on domain-independence, scalability and automation made the task of cause-effect identification a prime candidate for various machine learning approaches. In [19] authors proposed a semi-supervised method to automatically identify linguistic patterns that indicate causal relations in text. This work advocated the use of WordNet hierarchical classes like human action, phenomenon, state, psychological feature and event, as distinguishing features. The authors focused on using explicit patterns of the form $\langle$*NounPhrase1 - CausativeVerb - NounPhrase2* $\rangle$. On the other hand, authors of [7] and [5] suggested the use of causatives only. They used a hierarchical organization of several generic semantic templates. Bui et al. presented a novel method exploiting rules to extract and combine relationships between HIV drugs and mutations in viral genomes from articles in [16]. Girju et al.[20] addressed the problem as one of automatic recognition of relations between pairs of nominals in a sentence. Bui et al. [16] employed logistic regression to extract drugs (cause) and virus mutation (effect) occurrences from medical literature. Sorgente et al. [1] combined rule based and machine learning methods to take advantage of both. They used logical rules based on the dependency between words to extract possible cause-effect pairs, after which they applied Bayesian inference to reduce the number of pairs produced by ambiguous patterns. The relatively untouched task of extracting implicit cause-effect from sentences was tackled by Ittoo et al. [21]. Radinsky et al. used statistical inference mechanisms combined with hierarchical clustering to predict future events from news [3].

In SemEval-2007, a relation extraction task was proposed, that included 7 different kinds of relations, of which Cause-Effect was one type [22]. In 2008 Beamer et al.[23] formulated the 2007 SemEval task4 as the task of automatic classification of semantic relations between nouns. They proposed a WordNet-based learning model which relies on the semantic information of the constituent nouns. In 2009 also a similar task was proposed [24], but it was extended to classify causal relation between a pair of phrases in a sentence. One of the most used annotated datasets available for causal relation extraction comes from SemEval-2010 Task 8. However, other than a few exceptions, most of annotations here are over single-word (noun) cause and effect components.

Use of deep neural architectures to detect causal relations started thereafter. Most of the papers report their results for the SEMEVAL-2010 dataset, along with additional datasets created by them to increase the volume of training and test data. Xu et al. employed a combination of shortest dependency paths (SDP) and LSTMs to solve the relation classification task [25]. The shortest dependency path between two entities was identified and then classified using a neural architecture. Along with SDP, words and their parts-of-speech(POS) tags, grammatical relations and WordNet hypernyms were also used. In [8], Zhao et al. worked on predicting future events based on causal relations expressed in text. This work proposed using a new Restricted Hidden Naive Bayes model to extract causality from texts, using word-baed features, contextual features, syntactic features, position features and also causal connectives. In [26], Dasgupta et al. proposed a linguistically informed recursive neural network architecture for automatic extraction of cause-effect relations from text. Beside using content embedding, the BiLSTM based model also exploited other linguistic features.

In 2019, Yu et al. explored the use of causal language to report scientific findings. In [27] they proposed a BERT-based prediction model to classify sentences into four categories - “no relationship”, “correlational”, “conditional causal”, and “direct causal”. They reported achieving an accuracy of 0.90 and a macro-F1 of 0.88. Their results were obtained on a corpus of 3,000 PubMed research conclusion sentences, developed by them. Li et al. [28] proposed the SCITE architecture, which uses Self-attentive BiLSTM-CRF wIth Transferred Embeddings to extract cause and effect sequences directly without extracting candidate causal pairs and identifying their relations separately. The task was interpreted as a sequence tagging task where the text sequence is given as input and the model predicts which word in the output sequence is beginning or end of cause and effect phrases. Their proposed architecture also reports F-measure of 0.89 and 0.90 for cause and effect sequences on an extension of the SEMEVAL 2010 dataset. In a recently published paper, Huminski et al. have employed a rule-based method for automatic extraction of causal chains from text [29]. Extensive survey on extraction of causal relations from natural language text can be found in [6], and in the recently published work by Yang et al. in [30].

The following works have looked at different applications of causality detection and its applications. Egami et al. introduced a conceptual framework for making causal inferences with discovered measures as a treatment or outcome [31]. In their position paper [32], Blomqvist et al. highlight the utility of causal relation detection from biomedical text to distinguish causation from correlation. They propose the use of BERT [9], a transformer based language model, to extract causal relations from text, and store them in in the form of a Knowledge Graphs (KG), after being refereed by experts. Guo et al. have also used an unsupervised learning model to extract causal relations between pressure injury and risk factors [33], to construct a causal graph. In [34], Veitch et al. explored the possibility of using causality from text to understand what affects a scientific paper's acceptance. They presented methods to estimate causal effects from observational text data, adjusting for confounding features of the text such as the subject or writing quality. Keith et al. and Weld et al. have explored the possibility of using text data as a source of information to deal with potential confounders while doing causal reasoning in [35] and [36].

To summarize, it can be seen that most of the research work have worked with the assumption that causal sentences are given and the task is to isolate the cause and effect components. Others have used pattern based detection of causal sentences and thereafter isolated the cause-effect components. However, in reality, for any application, it cannot be expected that causal sentences will be detected separately and fed to a model for cause-effect identification. Our intent is to address this issue with joint modeling of the two tasks.

3. Proposed Architecture for Joint Modeling for Causality Detection and Cause-effect Extraction

In this section we present the details of the proposed joint model for causal sentence classification and causal relation extraction based on an enhanced architecture that is built on top of BERT-based language model, presented by Devlin et al. in [9]. The BERT model generates contextual embedding of the input text, which is thereafter fed to a CNNBiLSTM layer followed by a fully connected layer that jointly perform the sentence classification and sequence labeling tasks. Figure 1 presents the proposed architecture.

Causal sentence detection is modeled as a binary classification problem. The predicted label $y^1 \in \{0, 1\}$, where 0 stands for a non-causal sentence and 1 indicates that the sentence contains causal relations. The cause-effect relation extraction is modeled as a sequence labeling task that tags the input word sequence $x = (x_1, x_2, ..., x_T)$ with the label sequence $y^s = (y_1^s, y_2^s, ..., y_T^s)$ where the label $y_i^s \in$ *cause(C), effect(E), causal connective(CC) and None(N)*. For example, the labeled output of the sentence *The minister stated that gasoline is up because of refinery issues in Texas* will be *The/N minister/N stated/N that/N Gasoline/E is/E up/E because/CC of/CC refinery/C issues/C in/C Texas/C*. Our intent is to generate correct labels for an entire sequence of words that comprise cause, effect or causal connectives in a sentence. The sequence label prediction of a single word is dependent on predictions for surrounding words. It has been shown that structured prediction models such as cascaded CNN and BiLSTM models can significantly improve the sequence labeling performance. In this work, we have exploited the efficacy of a CNNBiLSTM layer by adding it on top of the BERT model. While the transformer model of BERT creates contextual representation of each word as well as the whole sentence, the spatial collocation properties commonly observed among a set of words bound by a specific relation, are captured by the CNNBiLSTM layer. While analyzing the results we observe that the joint model performs better than single task models as it learns the intricacies of causal dependencies among words and clauses and thereby improves the performance of both the tasks effectively. While more detailed results on standard datasets are presented later, here we present a few unique examples to highlight how the proposed model handles some complex situations correctly :

(1) - *The tornado destruction is followed by widespread disease.* - the model correctly identifies this as a causal sentence, where a causal relation is indicated by the words *is followed by*.
(2) *The leader was followed by his supporters in the march.* - the model correctly identifies that this is not a causal sentence despite the presence of the word *followed by*.
(3) *Climate change due to trapping of Green house gases, threatens people with food and water scarcity, increased flooding, extreme heat, more disease, and economic loss.* - Given this complex sentence without any known causal connectives, the model can not only detect it as causal but also identifies *climate change* as cause, and *threatens people with food and water scarcity, increased flooding, extreme heat, more disease, and economic loss* as effect.
(4) *Human migration and conflict can be a result* - this sentence is identified correctly by the model as non-causal, yet it could identify *conflict* as one of the effects.

Since we use BERT for language modeling, the proposed model is not restricted to work with sentences. It can work on collection of sentences as well. Due to this the model is capable of detecting implicit causalities between sentences also. In [37], Chen et al. had shown that joint modeling was found to outperform other models for the task of intent classification and slot-filling with NERs. Our work also reaffirms this.

3.1. Learning the Joint Model for Sentence Classification and Sequence Labeling Tasks

To solve the causality classification and causal relation extraction tasks together, we have trained the *CNNBiLSTM* layers using a loss functions with two separate components. The cross-entropy loss function is used for the purpose. Given a set of training

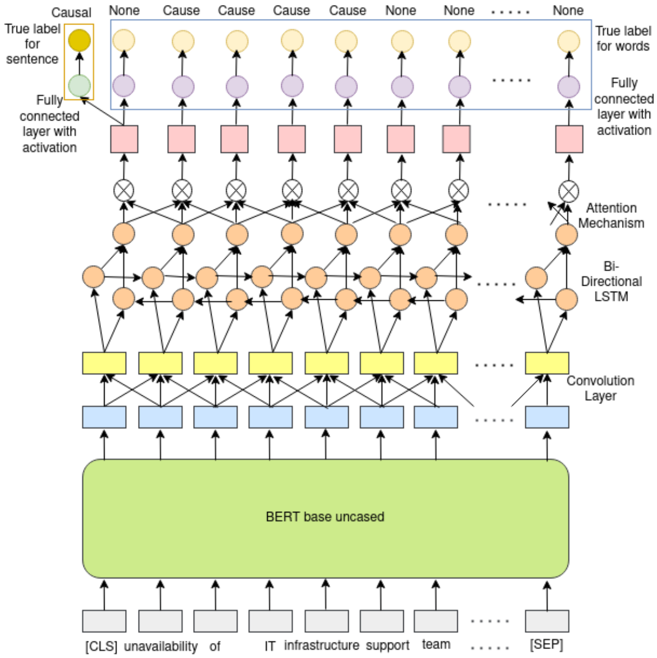

Figure 1. Overview of the neural network architecture for cause-effect relation extraction.

data x_t, w_t^i, $\bar{y}_t$ and $\bar{q}_t^i$, where x_t is the t-th word sequence to be predicted as *causal or not-a causal text*, w_t^i is the i-th word within the sentence to be predicted as *Cause (C), effect (E), causal connective (CN), and None(N)*, $\bar{y}_t$ and $\bar{q}_t^i$ are the one-hot representation of the ground-truth class labels for x_t and w_t^i respectively and y^i and y_n^s are the respective model predictions for both x_t and w_t^i. Thus, the causality classification is predicted as: $y^i = softmax(W_i * h_1 + b_i)$ On the other hand, for the sequence labeling task, we feed the final hidden states of the BERT-CNNBiLSTM network of the other tokens, $h_2, ..., h_T$, into a softmax layer to classify over the sequence labels. To make this procedure compatible with the Word-Piece tokenization, we feed each tokenized input word into a Word-Piece tokenizer and use the hidden state corresponding to the first sub-token as input to the CNNBiLSTM network and finally to a softmax classifier. The output of the model is represented as: $y_n^s = softmax(W^s * h_n + b_s), n \in (1...N)$ where h_n is the hidden state corresponding to the first sub-token of word x_n. Thus, the loss functions for the text classification (L_1) and causal relation extraction task (L_2) are separately defined as: $L_1(\theta) = -\sum_{k=1}^{K} \bar{y}_t^k log(y_t)$ $L_2(\theta) = -\sum_{t=1}^{N} \sum_{j=1}^{J} \bar{q}_t^{i,j} log(q_t^i)$ Where y_t is the vector representation of the predicted output of the model for the input sentence x_t. Similarly, q_t is the vector representation of the predicted output of the model for the input word w_t^i. K and J are the number of class labels for each task. N is the number of words of a given sentence for which the sequence labeling task is performed. The model is fine-tuned end-to-end via minimizing the cross-entropy loss.

We now define the joint loss function using a linear combination of the loss functions of the two tasks as:

$$L_{joint}(\theta) = \lambda * L_1(\theta) + (1 - \lambda) * I_{[y_{sentence}==1]} * L_2(\theta) \qquad (1)$$

Where, λ controls the contribution of losses of the individual tasks in the overall joint loss. $I_{[y_{sentence}==1]}$ is an indicator function which activates the causal relation labeling loss only when the corresponding sentence classification label is 1, since we do not want to back-propagate relation labeling loss when the corresponding sentence classification label is 0.

4. Data Description

Though causal relation extraction is acknowledged as an important NLP task, there is not much data available for training or evaluating models. Along with the other datasets mentioned in section 3, a new comprehensive dataset was released for a challenge called Cause-Effect Relation EXtraction (CEREX) during FIRE 2020 [2]. The CEREX-2020 dataset includes the causal dataset from SemEval-2010 task8 and extends it with more causal data from multiple sources. For the present work, we have also added 3000 causal sentences curated from Project Analyst reports available within the organization. Details about the CEREX dataset is given below:

1. **BBC News Article dataset** - this was created by the Trinity College Computer Science Department, from 140 News articles in five topical areas : business, sports, tech, entertainment and politics from 2004-2005 [38]. Out of the 1950 sentences in this collection, around 500 sentences were found to contain causation.
2. **SemEval-2010*:** This is a modified annotation of the original SemEval 2010 Task 8 data, released at CEREX-2020. The SemEval 2010 Task 8 data set proposes a entity based relation extraction task. Altogether there were 4300 sentences out of which only 1331 sentences were found to be having causal relations. While originally, only the entities were annotated as cause or effect, CEREX-2020 provided enhanced annotations for entire clauses. For example, in the sentence *The addition of water to the tank caused a runaway chemical reaction*, the SemEval 2010 annotated cause was “water” while the effect was “chemical reaction”. In CEREX-2020, the cause annotation was extended to the complete phrase “The addition of water to the tank” and the effect annotation is “runway chemical reaction”.
3. **The adverse drug effect (ADE) dataset** - This was released by [39]. It consists of information about consumption of different drugs and their associated side effects.
4. **Vehicle Recall News** - contains a collection of 1050 sentences collected from News articles that reported recall of different vehicles all over the world, along with the reasons for recall.

It is to be noted that all the above datasets contained only causal sentences, wherein each sentence contained at least one cause or an effect. Additionally, we also collected 10,000 more sentences collectively from the above domains, selecting them carefully so that they do not have any causal information. Table 1 presents brief statistics about the entire dataset used for the task.

[2] https://sites.google.com/view/cerex-fire2020/home

No.	**Source**	**Sentence count**	Part of CEREX-2020
1.	Project Analyst Report (AR)	3000	No
2.	BBC News(BBC)	503	Yes
3.	SEMEVAL* (SEM)	1331	Yes
4.	Adverse Drug Effect (ADE)	3821	Yes
5.	Recall News (VR)	1126	Yes
6.	Non-causal Sentences	10,000	Yes

Table 1. Data Statistics

5. Experiment and Results

We have designed two different experiments with the datasets described above. The experiments are distinguished from each other by the way, data from different sets are used for training, development and testing purposes.

Experiment-I (Training and Testing on Individual Dataset):, Each of the five data sets are used separately for evaluating the model's performance. Each dataset is randomly divided into 80%, 10% and 10% for training, testing and development respectively. The results are evaluated using five-fold cross-validation. Since, the 10% testing data in fold-1 is different from the 10% testing data in fold-2 or fold-3, we have reported the average performance. While working with the ADE dataset only, we used Bio-BERT, which is another base model, specifically trained on Bio-medical literature, instead of BERT-base.

Experiment-II (Training and Testing on Mixed Datasets): Data from the five sets are combined together and divided into training, development and test sets in the ratio of 80%, 10% and 10% respectively.

The pre-trained BERT-base model, presented in [40], uses 12 layers of transformer blocks, 768 hidden units, and 12 self-attention heads to provide a powerful context-dependent sentence representation. Most of the time, the same is used to fine-tune for the target causality detection tasks. For all the tasks, we have used Xavier initialization [41] for faster convergence. Further, we set the early stopping of fine-tuning to 800 steps in order to prevent over-fitting. We use a batch size 32, a maximum sequence length of 128, and a learning rate of $2 * 10^5$ for fine-tuning this model.

5.1. Baselines

The performance of the proposed joint model was compared with the following baseline models using the aforementioned datasets.

- The causal sentence classification and cause effect relation extraction model that was proposed in [42] had adopted a neural machine translation (**CEREX_NMT**) based architecture and reported results for CEREX dataset.
- A linguistic feature based transfer learning (**CEREX_FTL**) framework for Cause-Effect Relation Extraction was proposed in [43].
- A linguistically informed BiLSTM (**Li-BiLSTM**) model was proposed in [26], which uses word embeddings along with linguistic features.
- Self-attentive BiLSTM-CRF **(SCITE)** model, proposed in [28], employ different deep learning modelsto extract cause and effect components. This paper reported results for the SemEval-2010 Task 8 dataset. We took the openly available model, further trained it over the CEREX dataset and have presented the test data results in table 3. Since, this model does not implement detection of causal sentences, so it is considered only for Task 2.

Most of the models discussed above uses standard pre-trained word embeddings like, Glove, Word2Vec and Flair. Moreover, most of these models are trained to either classify sentences as causal/non-causal or to extract the respective cause-effect relations, but not perform both simultaneously.

In addition to the LSTM-based models, we also used a number of BERT based language models as our baseline, to evaluate the effect of the joint cost function, as well as contribution of the added CNNBiLSTM layers on top of BERT.

- **Single Task** $BERT_{base}$: This is the vanilla BERT model that has been specifically fine-tuned for two independent tasks to be conducted one after another. The first task is to perform causal sentence classification and then further trained to perform the sequence labeling task for causal relation extraction. This doesn't use CNN or BiLSTMs.
- **Single task BERT-CNNBiLSTM**: Here, in addition to the above vanilla BERT models, we have added a layer of CNN and an attention based BiLSTM units to both the units.
- **Joint** $BERT_{base}$: This is similar to the proposed joint model for sentence classification and sequence labeling tasks, however, without the top CNNBiLSTM layer.

We have performed a number of experiments to evaluate and compare the performance of our proposed system with the models presented above.

5.2. Results

Table 2 presents the results for task 1, i.e. causal sentence detection for Experiment-I. It is clear that the Joint BERT-CNN-BiLSTM model performs best as compared to the BiLSTM based neural networks as well as BERT based single task architectures. The joint model shows an average improvement of around 0.08 for F-score, as compared to the baseline single task BERT-CNN-BiLSTM model. Further, it is seen that the proposed model achieves high individual F-scores for all the datasets. For the SEMEVAL dataset, the joint BERT-Base classifier performs best with F1 score 0.95. Overall, it is obvious that joint task modeling definitely helps.

Table 3 presents the results for Experiment-1 of task 2 i.e. cause-effect relation extraction. Once again, the proposed joint BERT-CNN-BiLSTM model achieves the best F-measures of 0.82, 0.84, 0.94, and 0.88 for project analyst reports, BBC News, SEMEVAL and Recall news respectively. For SEMEVAL dataset, the earlier highest F1 measures reported for cause and effect was 0.89 and 0.90 by [27]. Having achieved F1 measure of 0.94 for causes, the proposed model clearly outperforms them in finding cause sequences, while remaining comparable in detecting effects. It was observed that the proposed joint BERT-CNNBiLSTM model significantly reduces the false negative scores and achieves a high true positive score, thereby achieving a higher F-measure as compared to the baseline single task models.

Table 4 presents the results for Experiment II, in which 80% of the entire combined dataset was used for training and the remaining 20% for testing. The aim was to see whether using a large dataset helps. Since joint BERT-CNNBiLSTM performed best so we report the results for only this model here. The first row shows the overall performance of the model, while the subsequent rows show break-up of the results for both the tasks. Though we see a general dip in the performance, F1 score for classification of

Models	**Dataset**	**Precision**	**Recall**	**F1-Score**
CEREX_NMT [42]	Analyst Reports(AR)	-	-	-
	BBC News (BBC)	0.31± 0.04	0.42± 0.03	0.35± 0.08
	SEMEVAL*(SEM)	0.39± 0.11	0.47± 0.14	0.41± 0.01
	Adverse Drug(ADE)	0.23± 0.07	0.36± 0.03	0.25± 0.06
	Recall News (VR)	0.37± 0.02	0.43± 0.14	0.40± 0.10
CEREX_FTL [43]	Analyst Reports(AR)	-	-	-
	BBC News (BBC)	0.61± 0.14	0.72± 0.03	0.65± 0.08
	SEMEVAL*(SEM)	0.69± 0.11	0.77± 0.14	0.71± 0.01
	Adverse Drug(ADE)	0.63± 0.07	0.76± 0.03	0.65± 0.06
	Recall News (VR)	0.77± 0.02	0.83± 0.14	0.80± 0.10
Li-BiLSTM [26]	Analyst Reports(AR)	0.71 ± 0.11	0.87± 0.09	0.78 ± 0.11
	BBC News (BBC)	0.71± 0.14	0.82± 0.03	0.75± 0.08
	SEMEVAL*(SEM)	0.79± 0.11	0.87± 0.14	0.81± 0.01
	Adverse Drug(ADE)	0.73± 0.07	0.86± 0.03	0.75± 0.06
	Recall News (VR)	0.87± 0.02	0.93± 0.14	0.87± 0.10
Single Task BERT-CNNBiLSTM	Analyst Reports(AR)	0.84± 0.11	0.87± 0.04	0.82± 0.03
	BBC News (BBC)	0.81± 0.01	0.82± 0.03	0.79± 0.04
	SEMEVAL*(SEM)	0.89± 0.03	0.87± 0.01	0.86± 0.05
	Adverse Drug(ADE)	0.83± 0.07	0.96± 0.02	0.90± 0.05
	Recall News (VR)	0.87± 0.08	0.93± 0.05	0.89± 0.06
Joint Task J-BERT	Analyst Report	0.87± 0.14	0.92± 0.09	0.85± 0.07
	BBC News (BBC)	0.81± 0.01	0.92± 0.05	0.84± 0.06
	SEMEVAL*(SEM)	0.81± 0.05	0.97± 0.03	0.89± 0.02
	Adverse Drug(ADE)	0.91± 0.03	0.96± 0.06	0.93± 0.05
	Recall News (VR)	0.87± 0.04	0.93± 0.05	0.90± 0.07
Joint Task J-BERT+CNNBiLSTM	Analyst Report	0.91± 0.05	0.97± 0.07	**0.90± 0.02**
	BBC News (BBC)	0.94± 0.13	0.97± 0.16	**0.94± 0.09**
	SEMEVAL*(SEM)	0.91± 0.15	0.97± 0.07	**0.94± 0.05**
	Adverse Drug (ADE)	0.89± 0.05	0.97± 0.01	**0.97± 0.05**
	Recall News (VR)	0.89± 0.02	0.95± 0.01	**0.92± 0.02**

Table 2. Reported results for Task-1: Causal sentence classification. Results are depicted for experiment-I, reported in terms of Precision, Recall and F1-Scores (along with the standard deviations). *Note that the baseline SCITE [28] model does not perform causal relation classification task, hence is not considered for the Task-1 evaluation. Further, the CEREX_NMT and CEREX_FTL models predate the *analyst report* dataset, and the models are not available, so results for this dataset could not be provided.

causal connectives have gone up significantly for most of the sets, other than for ADE. This can be explained as follows. The larger dataset exposed the model to a large variety of causal connectives, mostly coming from sentences which were a part of NEWS articles, and contributed to the overall gain. The ADE dataset uses terms specific to biomedical domain to indicate causality, hence showed no improvement. The cause and effect parts for each domain mostly contain domain-specific nouns and entities - hence these fields also did not show any gains. Causal sentence classification shows significant improvement for the Analyst project reports, while for others it remains same. This gain can possibly be attributed to the improved knowledge about causal connectors from the larger dataset. The only domain which loses is ADE. Given that it had attained a high F1 score of 0.97 for causal sentence classification, with BioBERT as the language modeler, our overall conclusion is that augmenting the language model with domain nuances is important for specialized domains.

Deeper dive into the results reveal that in around 12% cases the proposed model incorrectly predicted a cause or an effect event, whereas in only 5% of the sentences, the model incorrectly identified as “not a cause/effect” despite being marked as “cause/effect” by the experts. Sentences in the second category were found to contain ambiguous causal connectives such as: *from, by, based on the fact that* etc. The model can be further trained to recognize implicit causalities in phrases like *share-price manipulation scandal*, which contains a cause *manipulation of share price* and an effect *scandal*, without any causal connector between them.

Models	Dataset	Cause (C)	Effect (E)	Connectives (CC)
CEREX_NMT [42]	AR	-	-	-
	BBC	0.25± 0.10	0.28± 0.04	0.27± 0.02
	SEM*	0.31± 0.02	0.33± 0.07	0.35± 0.12
	ADE	0.29± 0.22	0.33± 0.01	0.31± 0.06
	VR	0.36± 0.03	0.42± 0.12	0.40± 0.02
CEREX_FTL	AR	-	-	-
	BBC	0.29± 0.11	0.36± 0.15	0.31± 0.12
	SEM*	0.33± 0.16	0.35± 0.31	0.33± 0.22
	ADE	0.39± 0.14	0.46± 0.22	0.41± 0.27
	VR	0.41± 0.07	0.43± 0.08	0.41± 0.02
Li-BiLSTM [26]	AR	0.73± 0.02	0.72± 0.02	0.61± 0.02
	BBC	0.68± 0.02	0.69± 0.02	0.74± 0.02
	SEM*	0.87± 0.02	0.83± 0.02	0.77± 0.02
	ADE	0.69± 0.02	0.74± 0.02	0.77± 0.02
	VR	0.80± 0.02	0.80± 0.02	0.76± 0.02
SCITE [28]	AR	0.75 ± 0.02	0.79± 0.14	0.77 ± 0.09
	BBC	0.71 ±0.07	0.76± 0.34	0.71 ± 0.21
	SEM*	0.78 ±0.10	0.79 ±0.20	0.80 ± 0.18
	ADE	0.81 ± 0.18	0.88 ±0.10	0.81 ± 0.05
	VR	0.86 ± 0.16	0.87 ±0.16	0.86 ± 0.07
Single Task BERT-CNNBiLSTM	AR	0.78 ±0.02	0.80± 0.15	0.78 ± 0.20
	BBC	0.78 ±0.12	0.77± 0.03	0.74 ± 0.19
	SEM*	0.81 ±0.23	0.86 ±0.06	0.83 ± 0.12
	ADE	0.89 ± 0.50	0.90 ± 0.01	0.82 ± 0.13
	VR	0.84 ± 0.13	0.88 ±0.03	0.90 ± 0.08
Joint Task BERT	AR	0.79 ±0.10	0.85± 0.13	0.79 ± 0.10
	BBC	0.80 ±0.18	0.78± 0.91	0.72 ± 0.02
	SEM*	0.91 ±0.02	0.88 ±0.05	0.84 ± 0.07
	ADE	0.94 ±0.80	0.93 ±0.16	0.86 ± 0.02
	VR	0.85 ± 0.08	0.88 ±0.16	0.92± 0.01
Joint Task J-BERT+CNNBiLSTM	AR	**0.87** ± **0.09**	**0.88** ± **0.10**	**0.89** ± **0.04**
	BBC	**0.84** ± **0.13**	**0.89**± 0.19	**0.79** ± **0.22**
	SEM*	**0.94** ± **0.22**	**0.88** ± **0.16**	**0.85** ± **0.09**
	ADE	**0.97** ± **0.13**	**0.97** ± **0.05**	**0.89** ± **0.01**
	VR	**0.88** ± **0.91**	**0.89** ± **0.16**	**0.94**± **0.03**

Table 3. Reported results for Task-II: Comparing F-scores of the Cause (C), Effect(E) and Connective (CC) extraction by the proposed Joint model with the baseline systems. The results are for Experiment-I.

	Task-1	**Task-2**		
Dataset	**Causal Sentence Classification**	**Cause**	**Effect**	**Connective**
Overall	**0.88**	**0.79**	**0.82**	**0.86**
AR	0.91	0.80	0.82	0.91
BBC	0.90	0.79	0.83	0.89
SEM*	0.92	0.81	0.86	0.87
ADE	0.80	0.75	0.76	0.72
VR	0.90	0.81	0.85	0.92

Table 4. Reported F1 scores for Experiment-II for the proposed Joint BERT-CNNBiLSTM for causal sentence classification (Task-1) and cause-effect extraction task (Task-2).

5.3. Extracting causal relations from a new unmarked text repository

We applied the joint BERT-CNNBiLSTM, with BioBERT as the language model, over a repository of 5000 article abstracts, taken from the Cord19 corpus released by a coalition of leading research groups and The White House to help in treatment of COVID 19 pandemic. Table 5 shows a few causal sentences along with the extracted components, each encapsulated within square brackets with the respective subscripts cause or effect. It can be seen that long clauses are correctly identified, irrespective of their positions, as shown in sentence 4. Additionally, the bio-medical entities drug(D), chemical(C), disease(I) and symptom(S), which were identified by the bio-medical Named Entity Recognizer (NER) called SciSpacy[44] are also marked in Table 5. Figure 2 shows a sample causal graph that is built from this output, where a causal relation is added between a drug or chemical and a disease or symptom, if the entities were part of a cause-effect pair in any sentence. Such causal graphs can be useful for drawing inference from large collections of text. This is still work in progress. We share the results to show that the model works well on

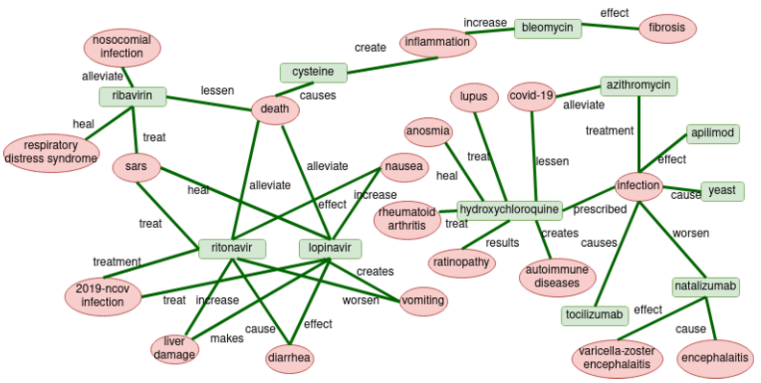

Figure 2. A partial network of a few representative clusters from the COVID-19 bio-medical journal dataset.

Sentences
1). We found that both [prophylactic and therapeutic $(Remdesivir)_{drug}$ had protective effects against Mers-cov replication and associated pathology]$_{cause}$, generally [resulting in]$_{connective}$ [less $(lung\ damage)_{symptom}$ and better $(pulmonary\ function)_{symptom}$]$_{effect}$ compared to controls.
2). The resulting [perturbation of the $(calcium)_{chemical}$ metabolism]$_{cause}$ can [cause]$_{connective}$ [various complications ranging from $(weakness)_{symptom}$ to $(osteoporosis)_{disease}$]$_{effect}$.
3). In addition , since they [received $(Lopinavir)_{drug}$/ $(Ritonavir)_{drug}$]$_{cause}$ which can also [cause]$_{connective}$ [$(diarrhea)_{disease}$]$_{effect}$, how much of gastrointestinal tract symptom was in fact related to Sars-Cov-2 or drug side effect?
4). [Severity of disease, viral replication, and $(lung\ damage)_{disease}$]$_{effect}$ were [reduced]$_{connective}$ when the [drug was administered either before or after infection with Mers-cov]$_{cause}$.
5). The [$(leukocytosis)_{disease}$ was unlikely to]$_{effect}$ be [caused by]$_{connective}$ [$(bacterial\ infection)_{disease}$]$_{effect}$ because [we excluded common bacteria or viruses associated with community acquired $(pneumonia)_{disease}$ and procalcitonin level of all patients in our study was no greater than 0.5 ng/ml.]$_{cause}$
6). [Combined usage with $(Ribavirin)_{drug}$]$_{cause}$ was also [associated with]$_{connective}$ [lower incidence of $(acute\ respiratory\ distress\ syndrome)_{disease}$, $(nosocomial\ infection)_{disease}$ and $(death)_{disease}$, amongst other favorable outcomes.]$_{effect}$
7). [$(Tocilizumab)_{drug}$ are not ideal]$_{cause}$ either as it can suppress the immune system and [lead to]$_{connective}$ [an increased risk of $(infection)_{disease}$]$_{effect}$.
8). [$(Remdesivir)_{drug}$ is a novel $(nucleotide)_{chemical}$ analog prodrug]$_{cause}$ that was intended to be used [for]$_{connective}$ the [treatment of $(ebola\ virus\ disease)_{disease}$]$_{effect}$.
9). [$(Diarrhea)_{drug}$, $(nausea)_{disease}$, $(vomiting)_{disease}$, $(liver\ damage)_{disease}$, and other adverse reactions]$_{effect}$ can occur following [combined therapy with $(Lopinavir/Ritonavir)_{drug}$]$_{cause}$.
10). Technically , we have little knowledge on the pathogen and pathogenesis , [without specific effectively drugs or vaccine against the $(virus\ infection)_{disease}$]$_{cause}$, which [cause]$_{connective}$ [difficulties in rescuing the severe cases]$_{effect}$ which [account for]$_{connective}$ [about 20 % of the $(infections)_{disease}$.]$_{effect}$

Table 5. Sample causes and effects mined from CORD19 corpus with Drugs, Chemicals, Diseases and Symptoms identified using SciSpacy

highly complex sentences from a completely unknown repository. The results are highly promising, though further improvements can be effected with more training.

6. Conclusion

In this paper, we present a multi-task learning architecture based on BERT enhanced with CNNBiLSTM layers to solve the dual problem of causal sentence classification and sequence labeling for cause, effect and connectives. The proposed model exploits the dependencies between the two tasks and improves the performance over baselines models. The performance of the joint model improves the state of the art results heretofore reported for SEMEVAL 2010 dataset. This work is now being extended to detect causal chains and implicit causalities. Chains need a different type of labeling scheme since a single clause may be an effect and a cause simultaneously. Current annotated datasets mostly contain simple sentences, with not too many implicit causalities. We are also extending this work to detect causal relations spread over multiple sentences. Annotation for the above tasks is under way. The other area we are working on is towards building causal knowledge graphs to be used for further reasoning, with humans in the loop. The first step however will be to build an indexed causal knowledge repository, which can be used for querying.

References

[1] Sorgente A, Vettigli G, Mele F. Automatic Extraction of Cause-Effect Relations in Natural Language Text. DART@ AI* IA. 2013;2013:37-48.

[2] Blanco E, Castell N, Moldovan D. Causal Relation Extraction. In: Lrec; 2008. .

[3] Radinsky K, Davidovich S, Markovitch S. Learning to predict from textual data. Journal of Artificial Intelligence Research. 2012;45:641-84.

[4] Girju R, Moldovan D. Mining answers for causation questions. In: AAAI symposium on mining answers from texts and knowledge bases; 2002. .

[5] Chan K, Low BT, Lam W, Lam KP. Extracting causation knowledge from natural language texts. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2002. p. 555-60.

[6] Asghar N. Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey. arXiv preprint arXiv:160507895. 2016.

[7] Low BT, Chan K, Choi LL, Chin MY, Lay SL. Semantic expectation-based causation knowledge extraction: A study on Hong Kong stock movement analysis. In: PAKDD; 2001. p. 114-23.

[8] Zhao S, Liu T, Zhao S, Chen Y, Nie JY. Event causality extraction based on connectives analysis. Neurocomputing. 2016;173:1943-50.

[9] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[10] Khoo CS, Kornfilt J, Oddy RN, Myaeng SH. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. Literary and Linguistic Computing. 1998;13(4):177-86.

[11] Do QX, Chan YS, Roth D. Minimally supervised event causality identification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2011. p. 294-303.

[12] Girju R. Automatic detection of causal relations for question answering. In: Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12. Association for Computational Linguistics; 2003. .

[13] Hobbs JR. Toward a useful concept of causality for lexical semantics. Journal of Semantics. 2005;22(2):181-209.

[14] Grishman R. Domain modeling for language analysis. DTIC Document; 1988.

[15] Garcia D. COATIS, an NLP system to locate expressions of actions connected by causality links. Knowledge acquisition, modeling and management. 1997.

[16] Bui QC, Nualláin BÓ, Boucher CA, Sloot PM. Extracting causal relations on HIV drug resistance from literature. BMC bioinformatics. 2010;11(1).

[17] Khoo CS. Automatic identification of causal relations in text and their use for improving precision in information retrieval; 1995.

[18] Khoo CS, Myaeng SH, Oddy RN. Using cause-effect relations in text to improve information retrieval precision. Information processing & management. 2001;37(1):119-45.

[19] Girju R, Moldovan DI, et al. Text mining for causal relations. In: FLAIRS Conference; 2002. p. 360-4.

[20] Girju R, Nakov P, Nastase V, Szpakowicz S, Turney P, Yuret D. Classification of semantic relations between nominals. Language Resources and Evaluation. 2009;43(2):105-21.

[21] Ittoo A, Bouma G. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In: International Conference on Application of Natural Language to Information Systems. Springer; 2011. p. 52-63.

[22] Girju R, Nakov P, Nastase V, Szpakowicz S, Turney P, Yuret D. Semeval-2007 task 04: Classification of semantic relations between nominals. In: Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics; 2007. p. 13-8.

[23] Beamer B, Rozovskaya A, Girju R. Automatic Semantic Relation Extraction with Multiple Boundary Generation. In: AAAI; 2008. p. 824-9.

[24] Hendrickx I, Kim SN, Kozareva Z, Nakov P, Ó Séaghdha D, Padó S, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Workshop on Semantic Evaluations: Recent Achievements and Future Directions. ACL; 2009. p. 94-9.

[25] Xu Y, Mou L, Li G, Chen Y, Peng H, Jin Z. Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 conference on empirical methods in natural language processing; 2015. p. 1785-94.

[26] Dasgupta T, Saha R, Dey L, Naskar A. Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks. In: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue; 2018. p. 306-16.

[27] Yu B, Li Y, Wang J. Detecting Causal Language Use in Science Findings. In: 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China; 2019. p. 4664-74.

[28] Li Z, Li Q, Zou X, Ren J. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. Neurocomputing. 2021;423:207-19.

[29] Huminski A, Bin NY. Automatic Extraction of Causal Chains from Text. LIBRES: Library & Information Science Research Electronic Journal. 2020;29(2).

[30] Yang J, Han SC, Poon J. A survey on extraction of causal relations from natural language text. arXiv preprint arXiv:210106426. 2021.

[31] Egami N, Fong CJ, Grimmer J, Roberts ME, Stewart BM. How to make causal inferences using texts. arXiv preprint arXiv:180202163. 2018.

[32] Blomqvist E, Alirezaie M, Santini M. Towards Causal Knowledge Graphs-Position Paper. In: Proceedings of Workshop on The Knowledge Discovery in Healthcare Data (KDH)@ ECAI; 2020. .

[33] Guo S, Jin L, Yang J, Jiang M, Han L, An N. Causal Extraction from the Literature of Pressure Injury and Risk Factors. In: International Conference on Knowledge Graph (ICKG). IEEE; 2020. p. 581-5.

[34] Veitch V, Sridhar D, Blei DM. Using text embeddings for causal inference. arXiv preprint arXiv:190512741. 2019.

[35] Keith KA, Jensen D, O'Connor B. Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. arXiv preprint arXiv:200500649. 2020.

[36] Weld G, West P, Glenski M, Arbour D, Rossi R, Althoff T. Adjusting for Confounders with Text: Challenges and an Empirical Evaluation Framework for Causal Inference. arXiv:200909961. 2020.

[37] Chen Q, Zhuo Z, Wang W. Bert for joint intent classification and slot filling. arXiv preprint arXiv:190210909. 2019.

[38] Greene D, Cunningham P. Practical solutions to the problem of diagonal dominance in kernel document clustering. In: ICML. ACM; 2006. p. 377-84.

[39] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of biomedical informatics. 2012;45(5):885-92.

[40] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems; 2017. p. 5998-6008.

[41] Jones A. An Explanation of Xavier Initialization. Retrieved from Andy's blog. 2015.

[42] Thenmozhi D, Arunima S, Amlan Sengupta AB. ssn_nlp@ FIRE2020: Automatic extraction of causal relations using deep learning and machine translation approaches. 2020.

[43] Aziz A, Sultana A, Hossain MA, Ayman N, Chy AN. Feature Fusion with Hand-crafted and Transfer Learning Embeddings for Cause-Effect Relation Extraction. In: FIRE (Working Notes); 2020. p. 756-64.

[44] Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:190207669. 2019.