Domain Adaptative Causality Encoder

Farhad Moghimifar[1], **Gholamreza Haffari**[2] and **Mahsa Baktashmotlagh**[1]
[1]The School of ITEE, The University of Queensland, Australia
[2]Faculty of Information Technology, Monash University, Australia
{f.moghimifar,m.baktashmotlagh}@uq.edu.au
gholamreza.haffari@monash.edu,

Abstract

Automated discovery of causal relationships from text is a challenging task. Current approaches which are mainly based on the extraction of low-level relations among individual events are limited by the shortage of publicly available labelled data. Therefore, the resulting models perform poorly when applied to a distributionally different domain for which labelled data did not exist at the time of training. To overcome this limitation, in this paper, we leverage the characteristics of dependency trees and adversarial learning to address the tasks of adaptive causality identification and localisation. The term adaptive is used since the training and test data come from two distributionally different datasets, which to the best of our knowledge, this work is the first to address. Moreover, we present a new causality dataset, namely MEDCAUS[1], which integrates all types of causality in the text. Our experiments on four different benchmark causality datasets demonstrate the superiority of our approach over the existing baselines, by up to 7% improvement, on the tasks of identification and localisation of the causal relations from the text.

1 Introduction

Causality is the basis for reasoning and decision making. While human-beings use this psychological tool to choreograph their environment into a mental model to act accordingly (Pearl and Mackenzie, 2018), the inability to identify causal relationships is one of the drawbacks of current Artificial Intelligence systems (Lake et al., 2015). The projection of causal relations in natural language enables machines to develop a better understanding of the surrounding context and helps downstream tasks such as question answering (Hassanzadeh et al., 2019), text summarisation (Ning et al., 2018),

and natural language inference (Roemmele et al., 2011).

The task of textual causality extraction can be divided into two main subtasks, causality identification and causality localisation. The former subtask focuses on identifying whether a sentence carries any causal information or not, which can be seen as classification problem. The objective of the latter subtask is to extract text spans related to cause and effect, subject to existence.

The automatic identification and localisation of causal relations in textual data is considered a non-trivial task (Dasgupta et al., 2018). Causal relations in text can be categorised as marked/unmarked and explicit/implicit (Blanco et al., 2008; Hendrickx et al., 2009). Marked causality refers to the case where a causal linguistic feature, such as "because of", is stated in the sentence. For example, in *"His OCD is because of genetic factors."*, *because of* is a causal marker, whereas in unmarked causality there is no such indicator. For instance, in *"Don't take these medications before driving. you might feel sleepy."* the cause and effect relationship is spread between two sentences without a marker. On the other hand, explicit causality refers to the case where both cause and effect are mentioned in text. However, in implicit causality, either cause or effect are directly mentioned in the text. In a more complex case called nested causality, multiple causal relations may exist in one sentence (e.g., *"Procaine can also cause allergic reactions causing individuals to have problems with breathing"*). All of these ambiguities contribute to the challenging nature of this task.

Traditional approaches to address the problem of causality extraction mainly relied on predefined linguistic patterns and rules to identify the existence of causal relations in a sentence (Mirza and Tonelli, 2016). More advanced approaches combined pattern-based methods with machine learning

---

[1]https://github.com/farhadmfar/ace

techniques (Zhao et al., 2018), and as such they require heavy manual feature engineering to perform reasonably. To overcome this problem, the recent approaches have adopted deep learning techniques to extract meaningful features from the text (Liang et al., 2019; Martínez-Cámara et al., 2017).

However, all the aforementioned approaches suffer from the problem of domain shift, where there is a distribution difference between the training and the test data. More specifically, the existing approaches perform poorly on the data from a new test domain (e.g. financial) which is contextually different from the training domain (e.g. medical).

To overcome the limitations of existing approaches on the tasks of causality identification and localisation, we propose a novel approach for domain-adaptive causality encoding which performs equally well when applied on the out-of-domain sentences. Our contribution is three-fold:

- To identify causal relationships and extract the corresponding causality information within a sentence, using graph convolutional networks, we propose a model which takes into account both syntactic and semantic dependency of words in a sentence. Extensive experimental results suggest that our proposed models for causality identification and localisation outperform the state-of-the-art results.

- We propose to use a gradient reversal approach to minimise the distribution shift between the training and test datasets. Our proposed adaptive approach improves the performance of the existing baselines by up to 7% on the tasks of adaptive causality identification.

- To fill the gap of the current causality datasets on encompassing different types of causality, we introduce MEDCAUS, a dataset of 15,000 labelled sentences, retrieved from medical articles. This dataset consists of sentences with labels of explicit, implicit, nested, and no causality.

2   Related Works

The projection of causal relation in textual data can be in various forms, depending on the type of causality. The categorisation mentioned in Section 1 can indicate the relation between pairs of events, phrases, concepts, named entities or a mixture of the aforementioned text spans (Hashimoto, 2019). Some works in the area have endeavoured to extract and present the textual information between concepts or events. Causal relations are a component of SemEval task (Hendrickx et al., 2009), but it involves a limited set of causal relations between pairs of nominals. Do et al. (2011) developed a framework based on combining semantic association and supervised causal discourse classification in order to identify causal relations between pairs of events. They expand the patterns in *pdtb* (Lin et al., 2009) using a self-training approach. Other methods (Riaz and Girju, 2014, 2013) leveraged linguistic features such as part-of-speech information, alongside with discourse markers, for identifying causal relations between events. An et al. (2019) used the syntactic patterns and word vectors to develop an unsupervised method for constructing causal graphs. To expand the repository of causal syntactic patterns, Hidey and McKeown (2016) built a parallel corpus between English Wikipedia and Simple Wikipedia, where the same causal relation might be in different syntactic markers in two parallel sentences. A supervised method was adapted by Mirza and Tonelli (2016) using lexical, semantic, and syntactic features within a sentence to address this task.

Using Hidey and McKeown (2016)'s method, Martínez-Cámara et al. (2017) created a set of labelled sentences, assuming all of the sentences include a causal relation, and presented a neural model based on LSTM to identify causality. Dasgupta et al. (2018) collected a dataset and developed a model using BiLSTM for extracting causal relation within a sentence[2]. Other approaches in event prediction applied Granger causality (Granger, 1988) to identify causal relations in time series of events (Kang et al., 2017). Rojas-Carulla et al. (2017) defined a proxy variable, which may carry some information about cause and effect, to identify causal relationship between static entities. Zhao et al. (2017) developed a causality network embedding for event prediction. De Silva et al. (2017) proposed a convolutional neural network model for identifying causality. Liang et al. (2019) also deployed a self-attentive neural model to address the task of causality identification, however, the extraction of causal information is not addressed by their model.

In more recent works, a dataset of counterfactual

---

[2]The source code and dataset are not publicly available.

sentences was released, as a part of SemEval2020 Task5 (Yang et al., 2020). The aim of this task is to identify and tag the existing counterfactual part of the sentence. Some works have attended to address this task using different deep learning architectures (Patil and Baths, 2020; Abi Akl et al., 2020). While counterfactuals are usually represented in form of a causal relation, this dataset does not cover different forms of textual causality.

As opposed to the aforementioned models, we propose a unified neural model for addressing both tasks of identifying and localising causality from a sentence. Our method leverages both syntactic and semantic relations within a sentence, and adapts to out-of-domain sentences.

3 Our Approach

In this section, we first describe the architecture of causality extractor, which uses graph convolutional networks (GCN) at its core. We then present how we make use of the adversarial learning strategy for adapting the model to new domains. Figure 1 illustrates the high level overview of our approach.

3.1 Graphical Causality Encoder (GCE)

Given a sentence $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, where $\boldsymbol{x}_i$ is the vector representation of the i-th token of the sentence, the goal of our model is two-fold: identifying whether or not the causal relation exists, and locating the position of cause and effect in the given sentence.

The core part of our causal identification model consists of an L-layer graph convolutional network (GCN) which takes as input the dependency tree of a sentence, obtained through Stanford CoreNLP (Manning et al., 2014). The dependency tree can be represented with an $n \times n$ adjacency matrix $\boldsymbol{A}$, where n is the number of nodes in the graph. In the adjacency matrix, $A_{ji} = A_{ij} = 1$ if an edge connects i to j, and zero otherwise. Given $\boldsymbol{h}_i^{(l-1)}$ as the representation of the node i at layer $l-1$, GCN updates the node representation at layer l as follows (Zhang et al., 2018; Kipf and Welling, 2016):

$$\boldsymbol{h}_i^{(l)} = \boldsymbol{f}^{\text{actv}}(\sum_{j=1}^{n} \tilde{A}_{ij} \mathbf{W}^{(l)} \boldsymbol{h}_j^{(l-1)} / d_i + \boldsymbol{b}^{(l)}) \quad (1)$$

where $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$, with $\boldsymbol{I}$ the identity matrix, $\boldsymbol{f}^{\text{actv}}$ an activation function (i.e., element-wise RELU), $\boldsymbol{b}^{(l)}$ the bias vector, $\mathbf{W}^{(l)}$ the weight matrix, and $d_i = \sum_{j=1}^{n} \tilde{A}_{ij}$ the degree of node i.

This formation captures the hidden embeddings of each token in a sentence with respect to its neighbours with maximum distance of L, with L the number of GCN layers. To take the words order and disambiguity into account and make the model less prone to errors from the dependency relations' results, we feed the word vectors into a bi-directional long short-term memory (BiLSTM) network. The output of the BiLSTM $\boldsymbol{h}_i^{(0)}$ is then used in the GCN, as illustrated in Equation 1. Hence, after applying the BiLSTM and GCN, each sentence is represented as:

$$\boldsymbol{h}_{\text{GCN}}(\boldsymbol{X}) = \boldsymbol{f}^{\text{pool}}(\text{GCN}(\text{BiLSTM}(\boldsymbol{X}))) \quad (2)$$

where $\boldsymbol{f}^{\text{pool}} : \mathbb{R}^{d \times n} \to \mathbb{R}^d$ is a pooling function generating the representations for the n tokens of the sentence. The final sentence representation is obtained by a feed forward network (FFNN) whose input is the concatenation of $\boldsymbol{h}_{\text{GCN}}(\boldsymbol{X})$ and $\boldsymbol{h}_{\text{BiLSTM}}(\boldsymbol{X})$. Note that $\boldsymbol{h}_{\text{BiLSTM}}(\boldsymbol{X})$ is the contextualised representation of the sentence from BiLSTM which is constructed by concatenating the leftmost and rightmost hidden states:

$$\boldsymbol{f}_{\boldsymbol{\theta}_{\text{enc}}}(\boldsymbol{X}) = \text{FFNN}([\boldsymbol{h}_{\text{GCN}}(\boldsymbol{X}); \boldsymbol{h}_{\text{BiLSTM}}(\boldsymbol{X})]) \quad (3)$$

where $\boldsymbol{\theta}_{\text{enc}}$ contains the collection of parameters of the GCN, BiLSTM, and the feed-forward network. This representation is then used to address the two main sub-tasks:

- For Task1, which is identifying causal relation within a sentence, we use this representation to get the probability of output classes,

$$P_{\boldsymbol{\theta}_{\text{class}}}(\text{causality}|\boldsymbol{X}) = \sigma(\boldsymbol{f}_{\boldsymbol{\theta}_{\text{enc}}}(\boldsymbol{X}).\mathbf{W}_{\text{class}} + b_{\text{class}}) \quad (4)$$

where $\boldsymbol{\theta}_{class} := \{\mathbf{W}_{\text{class}}, b_{\text{class}}\}$ contains the classifier's parameters, and σ is the sigmoid function.

- For Task2, locating cause and effect in a sentence, we use this representation to obtain the probability of the corresponding tag for each token. Since there are strong dependencies across tags, by adopting conditional random fields, we model the tagging decision jointly, with respect to surrounding tags. Consider $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]$ the sequence of tag predictions. The score corresponding to this sequence is defined as:

$$s(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=0}^{n} \boldsymbol{f}_{\boldsymbol{\theta}_{\text{enc}}}(\boldsymbol{X})_{i,\boldsymbol{y}_i} + \sum_{i=0}^{n} T_{\boldsymbol{y}_i, \boldsymbol{y}_{i+1}} \quad (5)$$
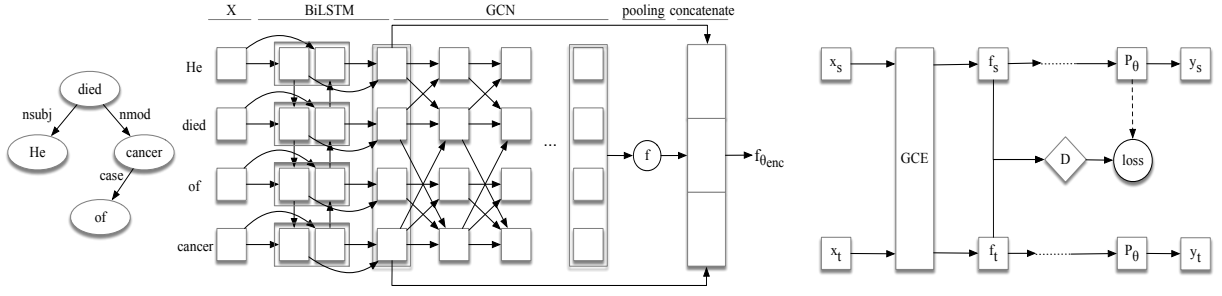
Figure 1: (Left) The dependency tree of the sentence *"He died of cancer."*. (Middle) Architecture of our proposed models for Causality Identifier (GCE), which uses the retrieved dependency tree as its core. (Right) The architecture of our proposed Adaptive Causality Encoder (ACE), which uses the structure of GCE as feature extractor.

where T is a square matrix with its size corresponding to the number of distinct tags. $T_{i,j}$ is representative of the score related to shifting from tag i to tag j. The probability of the tag sequence Y given X is then defined as (Y_X which denotes all possible sequences of tags for X):

$$P_{\theta_{\text{seq}}}(Y|X) = s(X, Y) - log(\sum_{y \in Y_X} e^{s(X,y)}) \quad (6)$$

Here, θ_{seq} contains the sequence tagger's parameters.

3.2 Adaptive Causality Encoder (ACE)

In this section, we represent a domain adversarial approach to adaptive causality identification and localisation. In unsupervised domain adaptation, we are given a source labelled data $D_s = \{(X_s^{(i)}, Y_s^{(i)})\}_{i=1}^{n_s}$ and unlabelled target data $D_t = \{X_t^{(j)}\}_{j=1}^{n_t}$. Our aim is to reduce the distributional shift between the two domains, and predict the labels of the target domain. Inspired by (Ganin et al., 2016; Long et al., 2018), we make use of an adversarial learning strategy, where the domain discriminator is trained to distinguish the source and domains, while the feature representation is trained to confuse the domain discriminator.

More formally, let us consider the following domain classifier,

$$P_{\theta_{\text{dom}}}(\text{source}|X) = \sigma(f_{\theta_{\text{enc}}}(X).\mathbf{W}_{\text{dom}} + b_{\text{dom}}) \quad (7)$$

where $\theta_{dom} := \{\mathbf{W}_{\text{dom}}, b_{\text{dom}}\}$ is the domain classifier's parameters. Our domain adversarial training objective is defined as,

$$\mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dom}}, \theta_{\text{task}}) := \sum_{(X,Y) \in D_s} \log P_{\theta_{\text{class}}}(Y|f_{\theta_{\text{enc}}}(X))$$
$$- \sum_{X \in D_s} \log P_{\theta_{\text{dom}}}(\text{source}|f_{\theta_{\text{enc}}}(X))$$
$$- \sum_{X \in D_t} \log \left(1 - P_{\theta_{\text{dom}}}(\text{source}|f_{\theta_{\text{enc}}}(X))\right).$$

The model parameters are then trained by,

$$\arg \max_{\theta_{\text{task}}} \max_{\theta_{\text{enc}}} \min_{\theta_{\text{dom}}} \mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dom}}, \theta_{\text{task}})$$

where minimization over θ_{dom} strengthens the accuracy of the domain classifier, but maximizing over θ_{enc} tries to confuse the domain classifier and strengthen the causality classifier.

3.3 Tagging Scheme

The objective of the task of causality localisation is to assign a label to each token in a sentence to locate the position of cause and effect. Cause and effect of a causal relation may span several tokens in a sentence. Therefore, the labels of a sentence usually are represented in the IOB-format (Inside, outside, and beginning). In this format, B-*label* indicates beginning of the span *label*, I-*label* shows a token inside the *label* but not the first token, and O-*label* represents the token as an outsider of *label*. However, inspired by (Ratinov and Roth, 2009) and (Dai et al., 2015), we use IOBES, an extended version of IOB, which also accounts for singleton labels and end of the label span token. Furthermore, to keep the tags consistent with the Equation 5, we add a *start* and *end* label to the set of tags.

4 Experiments

In this section, we first describe the datasets that have been used for the evaluation of our models, including our collected dataset. Then we present results of our proposed models on both (adaptive) causality identification and causality localisation.

4.1 Datasets

MEDCAUS We introduce our medical causality dataset with 15,000 sentences. The process of collection and annotation of the sentences was followed by the guideline of Hendrickx et al. (2009),

#causality classes	4
average sentence length	29.5
#explicit causality	9,092
#implicit causality	616
#nested causality	1,356
#non-causal	3,936
#Total sentences	15,000

Table 1: Statistics on MEDCAUS, our collected dataset.

including three main phases. In the first phase, sentences from medical articles of Wikipedia were randomly extracted. Using a wide variety of pre-defined causal connective words and patterns, we manually selected the sentences with potential causal relation and those without causal relation. In the second phase, the annotation instruction, multiple examples with different types of causal relation (i.e., explicit, implicit, nested, and non-causal) and different causal connective words were provided to the annotators. We asked four English-speaking graduate students to label the data accordingly. In the third phase, sentences with any disagreement that could not be resolved or were not clear in terms of causal relation were removed. To measure the level of agreement between our annotators, we give the same set of 1,000 sentences to the annotators. Using Fleiss Kappa measure (Fleiss and Cohen, 1973) ($\hbar$), the level of agreement between our annotators has been 0.71, showing the reliability of the annotations. Table 1 reports statistics about our collected dataset.

FinCausal The dataset, which is extracted from financial news provided by QWAM [3], includes different sets for both tasks of causality identification and localisation. For the former, it includes 22,058 sentences, and for the latter task, 1,750 sentences were provided[4].

SemEval-10 We use SemEval-10 Task 8, which has 1,003 sentences with causal relation. From other relations of this dataset, we randomly select 997 sentences, totalling 2,000 sentences. The sentences from this dataset are selected from a wide variety of domains, however, unlike MEDCAUS the causal relations are indicated only between pair nominals. This dataset was used for both causality identification and causality localisation task (Hendrickx et al., 2009).

BioCausal-Small The dataset is a part of larger dataset [5], consisting of 2,000 biomedical sentences from which 1,113 have causal relations. The sentences from this dataset have been collected from biomedical articles of PubMed [6]. Since this dataset only includes information about whether a sentence has causal relation (regardless of the position of the cause and effect), it has been used for causality identification (Kyriakakis et al., 2019).

4.2 Experimental Details

For both GCE and ACE, we use Stanford CoreNLP (Manning et al., 2014) to generate the dependency parsing tree for each sentence. We use the pre-trained 300-dimensional Glove vectors (Pennington et al., 2014) to initialise the embedding layer of our model. The hidden size for LSTM and the output feedforward layers is set to 100. We use the standard max pooling function for the pooling layer. Also, for all non-linearities in our model, we use Tanh function. A dropout ratio of $p = 0.5$ has been applied to all layers except for the last layer of GCN, for regularisation purposes.

For training of GCE, we split the data into train, development, and test set with the ratio of 60:20:20. For both models, we use batches of size 50. We train the model for 100 epochs, using Adamax optimiser. We use a decay rate of 0.9 if the $F1$ score of development set does not increase after each epoch. The reported results are micro-averaged precision, recall, and F1 score. All the hyperparameter and training settings were kept the same as reported above for other models for comparison. The original GCN and C-GCN model (Zhang et al., 2018), which have been used as baselines for experiment, use the Named Entity Recognition and Part of Speech Tagging embeddings of the related named entity as input to the model. Since identifying causal relation is not limited to named entities only, to be able to adjust baseline models to our experiment setup, we trained these model without the aforementioned embeddings.

4.3 Task1: Causality identification

In this section, we report the results on the task of identifying whether a sentence includes any causal relation or not. For this purpose, we use MEDCAUS and FinCausal to compare our GCE-based classifier (c.f. §3.1) with existing methods

		MEDCAUS			FinCausal	
Model	P	R	F1	P	R	F1
P-Wiki (Hidey and McKeown, 2016)	74.4	74.4	74.4	54.0	54.0	54.0
bi-LSTM (Martínez-Cámara et al., 2017)	84.2	**97.8**	90.5	81.3	77.0	79.1
GCN (Zhang et al., 2018)	90.8	94.6	92.7	85	74.8	79.6
C-GCN (Zhang et al., 2018)	91.2	94.9	93.0	**86.1**	68.7	76.4
GCE	**92.5**	94.0	**93.2**	84.8	**83.3**	**84**

Table 2: Results of our proposed method on Task1, causality identification, compared with the baseline approaches on the MEDCAUS and FinCausal dataset.

		MEDCAUS			FinCausal	
Model	P	R	F1	P	R	F1
bi-LSTM-CRF (Martínez-Cámara et al., 2017)	**77.4**	69.9	73.4	**82.4**	65.0	72.7
GCN-CRF (Zhang et al., 2018)	31.9	46.8	37.9	66.1	55.5	60.3
C-GCN-CRF (Zhang et al., 2018)	72.5	**75.9**	74.1	76.3	68.8	72.3
S-LSTM-CRF (Lample et al., 2016)	58.6	64.0	61.2	61.5	29.7	40.0
ELMO-CRF (Peters et al., 2018)	48.5	78.9	60.1	71.8	61.3	66.1
GCE	76.3	73.6	**74.9**	79.2	**69.8**	**74.2**

Table 3: Results of our proposed method on Task2, causality localisation, compared with the baseline approaches on the MEDCAUS and FinCausal dataset.

for causality identification. We divide the dataset into train/test/validation sets based on the ratio 60:20:20.

We compare our model to the P-Wiki (Hidey and McKeown, 2016), which is a rule-based method, and bi-LSTM (Martínez-Cámara et al., 2017). Furthermore, since this task is closely related to the task of relation extraction, we compare our model to GCN, and C-GCN (Zhang et al., 2018), which use dependency tree information of the sentence.

The results are reported in Table 2. Our GCE-based classifier achieves the highest F1 and precision score on MEDCAUS, amongst all the models, followed closely by C-GCN. However, bi-LSTM shows the highest recall score. On FinCausal, our proposed model achieves the highest F1 and recall score, comparatively, while C-GCN hits the highest score on precision. Given the complexity and ambiguity of the projection of causal relation in natural language, taking both semantic and syntactic relations of a sentence improves the model. Hence, as suggested by the results, using both contextualised representation of a sentence and dependency relations of tokens of a sentence enriches the model, and results in obtaining more accurate prediction of causal relations.

4.4 Task2: Causality Localisation

This section covers the results of the performance of our proposed model, compared to other models, in terms of extracting cause and effect from textual data. MEDCAUS and FinCausal are used in this task for evaluation purposes. Each dataset are split into train/test/validation with the ratio of 60:20:20. For comparison, we report the results of the performance of each model in labelling each token with the proper tag. For this purpose, precision, recall, and F1 score are reported.

Similar to the Task1, we compare our model to bi-LSTM (Martínez-Cámara et al., 2017), GCN, and C-GCN (Zhang et al., 2018). Also, we compare our model to the proposed model of Lample et al. (2016), with two variations of using S-LSTM and ELMO (Peters et al., 2018) for contextual embedding.

The results of causality localisation are reported in Table 3. The experiments on MEDCAUS show that while bi-LSTM-CRF achieves better results in precision, it fails to gain high recall. On the hand C-GCN-CRF achieves highest recall, followed closely by our model. However, in F1 score, our model, outperforms the baselines. On FinCausal, bi-LSTM-CRF achieves the highest precision. However, our model achieves better recall and F1 score.

4.5 Results of ACE

In this section, we present the results of our ACE model on the task of adaptive causality identification and causality localisation. To this end, we

	MEDCAUS → BioCausal			MEDCAUS → SemEval			MEDCAUS → FinCausal		
Models	P	R	F1	P	R	F1	P	R	F1
bi-LSTM (Martínez-Cámara et al., 2017)	76.1	57.6	66.0	82.5	62.8	71.3	47.6	8.3	14.2
GCN (Zhang et al., 2018)	75.4	51.0	60.8	78.4	67.6	72.6	49.2	53.2	51.1
C-GCN (Zhang et al., 2018)	71.3	42.9	55.3	84.1	70.8	76.9	48.7	52.3	50.4
bi-LSTM+DA	75.6	58.9	66.2	81.6	69.5	75.1	47.9	61.1	53.7
GCN+DA	72.8	70.3	71.5	82.9	66.7	73.9	46.8	57.4	51.6
C-GCN+DA	78.4	55.5	65.1	81.9	71.0	76.1	**49.1**	54.6	52.7
CDAN (Long et al., 2018)	**85.5**	50.1	63.8	**84.6**	73.8	78.8	43.6	53.3	48.0
CDAN-E (Long et al., 2018)	**83.8**	55.0	66.4	81.2	74.2	77.6	48.3	63.3	54.8
ACE	74.3	**77.1**	**76.7**	84.4	74.2	**79.0**	47.4	**74.0**	**57.8**

Table 4: Results on our proposed adaptive causality encoder compared with the baselines for the task of causality identifier. The source dataset is MEDCAUS. BioCausal, SemEval, and FinCausal are considered as the target dataset.

	MEDCAUS → SemEval			MEDCAUS → FinCausal		
Models	P	R	F1	P	R	F1
bi-LSTM (Martínez-Cámara et al., 2017)	16.3	52.2	24.9	**64.1**	16.1	25.8
GCN (Zhang et al., 2018)	8.8	29.6	13.5	41.6	40.9	41.0
C-GCN (Zhang et al., 2018)	18.9	47.6	27.1	63.8	13.0	21.6
bi-LSTM+DA	**51.2**	42.0	46.2	44.9	39.4	41.9
GCN+DA	9.1	25.5	13.4	39.6	45.2	42.2
C-GCN+DA	45.1	42.1	43.6	40.0	**45.5**	42.6
CDAN (Long et al., 2018)	40.9	49.7	44.8	36.9	42.8	39.6
CDAN-E (Long et al., 2018)	47.3	40.6	43.7	36.8	42.6	39.5
ACE	42.3	**53.6**	**47.3**	42.2	43.2	**42.7**

Table 5: Results on our proposed adaptive causality encoder compared with the baselines for the task of causality localisation. The source dataset is MEDCAUS and target dataset are SemEval and FinCausal.

consider MEDCAUS as the source domain, and SemEval-10, BioCausal, and FINCAUSAL as target domains [7]. We compare our model to bi-LSTM (Martínez-Cámara et al., 2017), GCN and C-GCN (Zhang et al., 2018) as the baselines, their domain adaptive versions (indicated with "+DA" in the tables), and a state-of-the-art approach of conditional adversarial domain adaptation (CDAN and CDAN-E) (Long et al., 2018).

Table 4 summarises the results of our experiments of domain adaptive causality identification. As it can be seen, adding the domain adaptive strategy to the baselines improves their performance on all target datasets, by up to 39%. Furthermore, while CDAN achieves a better precision, it fails to balance the recall and performs poorly in terms of F1 score. On the other hand, our model (ACE; c.f. §3.2), outperforms all of the other models in recall

and F1 score.

The results on applying domain adaptation method for the task of causality localisation is reported in Table 5. Applying our proposed domain adaptive model has improved the recall and F1 score of the baselines on both target datasets. While other models achieve better precision scores, our model consistently gains a better recall and F1 score, showing the superiority of our approach.

Visualisation Figure 2 visualises the effect of applying our proposed domain adaptation module (ACE; c.f. §3.2), to different target datasets. The extracted features ($f_{\theta_{enc}}$) of the source and target datasets are visualised using t-distributed Stochastic Neighbour Embedding (t-SNE) (Maaten and Hinton, 2008). The source and target datasets are shown in red and blue, respectively. In each subfigure, the features before and after applying ACE are represented on the left and right side, respectively. It is clear that, where the source and target domains data have different distributions, ACE

---

[7]Since BioCausal does not provide tags of cause and effect, this dataset was not used for domain adaptive causality localisation.

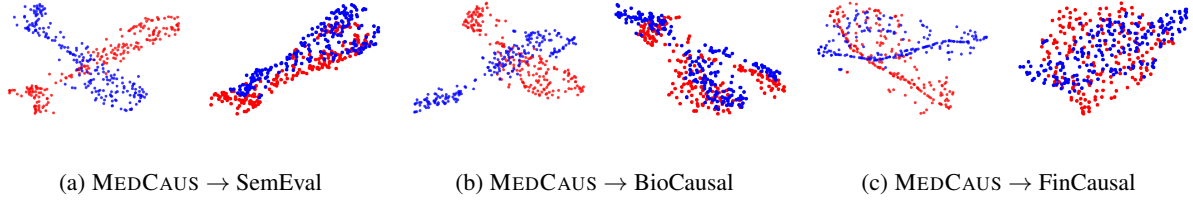|		
(a) MEDCAUS → SemEval	(b) MEDCAUS → BioCausal	(c) MEDCAUS → FinCausal

Figure 2: t-SNE visualisation of the domain adaptation task (with perplexity of 20 and Principal Decomposition Analysis (PCA) (Wold et al., 1987) initialisation). The source and target data are shown in red and blue, respectively. In each section, features before and after applying ACE are represented,on left and right side, respectively.

Causality Identification	MEDCAUS	1. Severe narrowings may cause chest pain (angina) or breathlessness during exercise or even at rest.	✓
		2. When the floor of the mouth is compressed, air is forced into the lungs.	✗
	SemEval	1. Mechanical faults caused delays and cancellations on Wellington's suburban train services this morning	✗
		2. The overall damage caused by the destruction of land and property for the Wall's construction has taken many years to recover further.	✓
	FinCausal	1. Thomas Cook, one of many world's largest journey corporations, was based in 1841 to function temperance day journeys, and now has annual gross sales of 39 billion.	✗
		2. The judge's decision converted the arbitration award to a legal judgement and the sum, including interest accrued since 2013, soared to more than $9 billion.	✓
	BioCausal	1. For cost and convenience reasons other altered fractionation schedules have been adopted in routine practice.	✗
		2. The sequential technique also minimises the incidence of iris bleeding.	✓
Causality Localisation	MEDCAUS	1. **A high rate of consumption** can also lead to cirrhosis, gastritis, gout, pancreatitis, hypertension, various forms of cancer, and numerous other illnesses.	
		2. **The phlegm produced by catarrh** may either discharge or cause a blockage that may become chronic.	
	SemEval	1. He took a shower after using hair cream to avoid skin irritation from the **chemicals in the product**.	
		2. A **cigarette** set off a smoke alarm.	
	FinCausal	1. The DGR in the Roth is lower at 5.4% due primarily to its **holding of REITs**.	
		2. Company tax receipts were $4.6 billion higher than predicted, mainly due to **mining profits**, but Mr Frydenberg could not say how much was due to strong iron ore demand.	

Table 6: Examples of the performance of our proposed model in causality identification and localisation. The top section of the table provides examples for the first task. The ✓ and ✗ indicate identification of causal and no causal relation in the sentence, respectively. The bottom part of the table presents examples for the second task. We have used underline with red colour and dashed underline with blue colour to show cause and effect respectively.

matches the distributions, which greatly helps with improving the performance on the target data.

4.6 Qualitative Analysis

In this section, we demonstrate the capability of our proposed models in addressing the tasks of causality identification and localisation. To this end, for each task, two sentences from each dataset are presented in Table 6. The top section of the table provides examples for causality identification. The bottom section presents example for causality localisation. The examples suggests that our proposed models perform accurately on datasets with different distributional features.

5 Conclusion

In this work, we propose a new dataset for the task of causal identification and causal extraction from natural language text. We further propose a neural-based model for textual causality identification and localisation, which makes use of dependency trees. We then make use of adversarial training to adapt the causality identification and localisation models to new domains. Empirical results show that our method outperforms state-of-the-art models and their adapted versions.

References

Hanna Abi Akl, Dominique Mariko, and Estelle Labidurie. 2020. Semeval-2020 task 5: Detecting counterfactuals by disambiguation. *arXiv*.

Ning An, Yongbo Xiao, Jing Yuan, Jiaoyun Yang, and Gil Alterovitz. 2019. Extracting causal relations from the literature with word vector mapping. *Computers in biology and medicine*.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*.

Hong-Jie Dai, Po-Ting Lai, Yung-Chun Chang, and Richard Tzong-Han Tsai. 2015. Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *Journal of cheminformatics*.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Annual SIGdial Meeting on Discourse and Dialogue*.

Tharini N De Silva, Xiao Zhibo, Zhao Rui, and Mao Kezhi. 2017. Causal relation identification using convolutional neural networks and knowledge based features. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Conference on Empirical Methods in Natural Language Processing*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*.

Clive WJ Granger. 1988. Some recent development in a concept of causality. *Journal of econometrics*.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from wikipedia. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. *IJCAI19*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Annual Meeting of the Association for Computational Linguistics*.

Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard Hovy. 2017. Detecting and explaining causes from text for a time series event. *arXiv preprint arXiv:1707.08852*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Manolis Kyriakakis, Ion Androutsopoulos, Artur Saudabayev, et al. 2019. Transfer learning for causal sentence detection. *arXiv preprint arXiv:1906.07544*.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Shining Liang, Wanli Zuo, Zhenkun Shi, and Sen Wang. 2019. A multi-level neural network for implicit causality detection in web texts. *arXiv preprint arXiv:1908.07822*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Conference on Empirical Methods in Natural Language Processing*.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Annual meeting of the association for computational linguistics: system demonstrations*.

Eugenio Martínez-Cámara, Vered Shwartz, Iryna Gurevych, and Ido Dagan. 2017. Neural disambiguation of causal lexical markers based on context. In *International Conference on Computational Semantics Short papers (IWCS)*.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *International Conference on Computational Linguistics: Technical Papers*.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Annual Meeting of the Association for Computational Linguistics*.

Rajaswa Patil and Veeky Baths. 2020. Cnrl at semeval-2020 task 5: Modelling causal reasoning in language with multi-head self-attention weights based counterfactual detection. *arXiv preprint arXiv:2006.00609*.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Conference on empirical methods in natural language processing (EMNLP)*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *SIGDIAL 2013 Conference*.

Mehwish Riaz and Roxana Girju. 2014. Recognizing causality in verb-noun pairs via noun and verb semantics. In *EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Mateo Rojas-Carulla, Marco Baroni, and David Lopez-Paz. 2017. Causal discovery using proxy variables. *arXiv preprint arXiv:1702.07306*.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*.

Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. 2020. Semeval-2020 task 5: Counterfactual recognition. In *International Workshop on Semantic Evaluation (SemEval-2020), Barcelona, Spain*.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Conference on Empirical Methods in Natural Language Processing*.

Sendong Zhao, Meng Jiang, Ming Liu, Bing Qin, and Ting Liu. 2018. Causaltriad: Toward pseudo causal relation discovery and hypotheses generation from medical text data. In *ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*.

Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Bin Wang, and ChengXiang Zhai. 2017. Constructing and embedding abstract event causality networks from text snippets. In *ACM International Conference on Web Search and Data Mining*.