Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks

Tirthankar Dasgupta, **Rupsa Saha**, **Lipika Dey** and **Abir Naskar**
TCS Innovation Lab, India
(dasgupta.tirthankar, rupsa.s, lipika.dey, abir.naskar)@tcs.com

Abstract

In this paper we have proposed a linguistically informed recursive neural network architecture for automatic extraction of cause-effect relations from text. These relations can be expressed in arbitrarily complex ways. The architecture uses word level embeddings and other linguistic features to detect causal events and their effects mentioned within a sentence. The extracted events and their relations are used to build a causal-graph after clustering and appropriate generalization, which is then used for predictive purposes. We have evaluated the performance of the proposed extraction model with respect to two baseline systems, one a rule-based classifier, and the other a conditional random field (CRF) based supervised model. We have also compared our results with related work reported in the past by other authors on SEMEVAL data set, and found that the proposed bi-directional LSTM model enhanced with an additional linguistic layer performs better. We have also worked extensively on creating new annotated datasets from publicly available data, which we are willing to share with the community.

1 Introduction

The concept of causality can be informally introduced as a relationship between two events e_1 and e_2 such that occurrence of e_1 results in the occurrence of e_2. Curating causal relations from text documents help in automatically building causal networks which can be used for predictive tasks. Expression of causality can be expressed within text documents in arbitrarily complex ways. For example, in the sentence "*Aircel files for bankruptcy over mounting financial troubles*", the event "*mounting financial troubles*" is causing the event "*Aircel filed for bankruptcy.*" In a more complicated scenario, *"Company recalled some vehicles to fix loose bolts that could lead to engine stall"* we can observe nested cause-effect pairs. Here, the effect "*company recalled vehicle*" is caused by the event "*to fix loose bolts is not easy to extract. That the cause "loose bolts" could lead to engine stall*", is even more difficult to detect.

While there has been a considerable body of researchers working in the area whose work has been reviewed in section 2, there are many challenges that are still not properly addressed. Most of the earlier approaches have considered rule based or traditional machine learning algorithms which heavily depend on careful feature engineering. Though one sees adoption of deep learning techniques for causality extraction, it is still considerably low compared to other text mining tasks. This is largely due to the unavailability of adequate annotated data: the only available dataset for evaluation is the SEMEVAL-10 Task 8 which is woefully inadequate to train such deep models. There are challenges with annotations of this data also (Rehbein and Ruppenhofer, 2017).

Most of the existing extraction mechanisms look for single word representation of events within a sentence, thereby yielding wrong results. For example, in the sentence *"The AIDS pandemic caused by the spread of HIV infection"* the cause and effect are both multi-word phrases i.e. "spread of HIV infection" and 'AIDS pandemic'. However, SEMEVAL 2010 annotated dataset for this task mentions the cause and effect as "infection" and "pandemic" only. In another example, "*Infectious diseases or communicable diseases are caused by bacteria, viruses, and parasites.*", the need to extract multiple causal as well as effect events is obvious. The example sentence in the first paragraph not only demonstrates the need to

Proceedings of the SIGDIAL 2018 Conference, pages 306–316,
Melbourne, Australia, 12-14 July 2018.

extract phrases as events, but also highlights how complex such statements can be, often without the use of known causal connectives like "*causes, because of, leads to, after, due to*" etc. which have been traditionally exploited by the community.

In this work, we explore the use of bidirectional LSTMs that can learn to detect causal instances from sentences. To address the paucity of training data, we propose the use of additional linguistic feature embeddings, over and above the regular word embeddings. With the use of such linguistically-informed deep architecture, we avoid the task of complex feature engineering.

A major contribution of this work is in developing annotated datasets with information curated from multiple sources spanning across different domains. To do this, we have collected news articles and generate annotations. Beside SEMEVAL dataset we have also used another available dataset that has annotated data about drugs and their adverse effect extracted from Medline (Gurulingappa et al., 2012). We have done intensive experimentations with parts of the dataset for training and testing which will be discussed in the following sections.

Detection of causal relation from text has many analytical and predictive applications. Few of these are: detecting cause-effect relations in medical documents, learning about after effects of natural disasters, learning causes for safety related incidents etc.. However to build a meaningful application that can detect an event from texts and predict its possible effects, there is a need to curate large volume of cause-effect event pairs. Further, similar events need to be grouped and generalized to super classes, over which the predictive framework can be built(Zhao et al., 2017). In this paper, we have proposed a k-means clustering of causal and effect events detected from text, using word vector representations.

The rest of the paper is organized as follows. Section 2 summarizes challenges and related works on causality detection. Section 3 presents the resource creation and the architecture of the proposed causality extraction framework. Experiments and evaluation are detailed in Section 4. Finally, in section 5 we conclude the paper.

2 Challenges in Causality Detection and the State of the Art

Identification of causality is not a trivial problem. Causation can occur in various forms. Two common differentiations are made on: a) *Marked and Unmarked causality* and b) *Implicit and Explicit causality* (Blanco et al., 2008)(Hendrickx et al., 2009)(Sorgente et al., 2013). Marked Causality is where there is a linguistic signal of causation present. For example, *"I attended the event because I was invited"*. Here, causality is marked by *because*. On the other hand in "*Drive slowly. There are potholes*", causality is unmarked.

Explicit Causality is where both cause and effect are stated. For example, "*The burst has been caused by water hammer pressure*" has both cause and effect stated explicitly. However, "*The car ran over his leg*" does not have the effect of the accident explicitly stated.

Automatic extraction of cause-effect relations are primarily based on three different approaches namely, Linguistic rule based, supervised and unsupervised machine learning approaches. Both SemEval-2007 (Girju et al., 2007) & 2010 (Hendrickx et al., 2009) had tasks aimed at identifying different relations from text, including Cause-Effect relations. Both tasks offered a corpus of annotated gold standard data to researchers. However, the task has primarily focused on extracting single word cause-effect pairs. Early work in this area relied totally on hand-coded patterns. These were heavily dependent on both domain and linguistic knowledge, due to the nature of the patterns, and were hard to scale up. PROTEUS (Grishman, 1988) and COATIS (Garcia, 1997) were two early systems that used such non-statistical techniques. C.G Khoo carried out extensive development of this train of thought in a series of works (Khoo et al., 1998) (Khoo et al., 2001), and eliminated a lot of the need for domain knowledge.

A method of automatically identifying linguistic patterns that indicate causal relations and a semi-supervised method of validation of patterns obtained was proposed by (Girju et al., 2002). In particular, this work introduced the usage of WordNet hierarchal classes, namely, human action, phenomenon, state, psychological feature and event, as a distinguishing feature.

Radinsky et al. in their work uses statistical inferencing combined with hierarchical clustering technique to predict future events from

news (Radinsky et al., 2012). Logistic regression was employed (Bui et al., 2010) to extract drugs (cause) and virus mutation (effect) occurrences from medical literature. The relatively untouched task of extracting implicit cause-effect from sentences was tackled by Ittoo et.al (Ittoo and Bouma, 2011). More recently, Zhao et al. (Zhao et al., 2017) have proposed novel causality network embeddings for the abstract representation of causal events from News headlines. Here, the authors have primarily used four common causal connectives namely, "because", "after", "because of" and "lead to" to extract causal mentions in news headlines and constructed a network of causal relations. The authors have proposed a novel generalization technique to represent "specific events" into more abstract form. Finally, they proposed a dual cause-effect model that uses the causal network embeddings and optimize the margin based loss function to predict effect of a given cause. Although the work is commendable, there are various factors that need to be addressed further. For example, construction of the causal network itself is a non trivial task. Some of the linguistic challenges have already mentioned earlier in this section. Further, Zhao et al. worked with only unambiguous causal connectives. On the contrary causal connectives can be ambiguous also (Sorgente et al., 2013) (Hendrickx et al., 2009) For example, *from* in *"Profits from the sale were given to charity"* implies causation of *profits* due to *the sale*, while *from* in *"Sales profits increased from 1.2% to 2%"* does not have any causality involved in it. Analysis of such complex constructs are yet to be addressed.

3 Proposed Methodology

The overall architecture of our proposed approach is composed of three modules: a)Resource Creation b) Linguistic preprocessor and feature extractor, c) Classification model builder, and d) Prediction framework for cause/effect, built on the output of the classifier module. Each of the individual modules are described in the following subsections.

3.1 Resource Creation

Data Description: In this section we will discuss about the following dataset used to develop and test our proposed models. 1) Part of the SemEval 2010 Task 8 data set dealing with"Cause-Effect" relation, which consists of 1331 sentences. 2) The adverse drug effect (ADE) dataset (Gurulingappa et al., 2012) composed of 1000 sentences consisting of information about consumption of different drugs and their associated side effects. 3)The BBC News Article dataset, created by the Trinity College Computer Science Department, containing news articles in five topical areas : business, sports, tech, entertainment and politics from 2004-2005 (Greene and Cunningham, 2006). We have considered 140 business news articles, containing approximately 1950 sentences. Out of this, around 500 sentences were found to contain causation. 4)Around 4500 analyst reports of a specific organization over a period of seven months is the fourth dataset that we have considered. We have manually extracted all the sentences that contained causation. 5) The Recall dataset [1] is a collection of 1050 recall news of different products.

Table 1: Data Statistics

Source	Sentence count	Avg. sent. length
Analyst Report (AR)	4500	23.7
SEMEVAL (SEM)	1331	18.7
BBC News(BBC)	503	22.5
ADE	3000	20.5
Recall News (RN)	1052	23.1

The first two datasets, that is, SemEval and ADE datasets, are already publicly available. However, for the SemEval dataset we have extended the annotation to phrase-level causal relationships. Hence the fresh annotations of these existing data sets, as well as parts of the annotated Recall news and BBC news datasets, will be publicly shared with this paper. We could not share the analyst report dataset due to copyright and IPR issues.

Preprocessing: We perform a number of preprocessing over the collected dataset. The first stage of preprocessing involves identifying which sentences are probably candidates for cause-effect identification out of a body of text. This involves looking for the presence of at least one causal connective in the sentence under consideration. Xuelan (Xuelan and Kennedy, 1992) reported a list of 130 causal connectives in English. To extend the list we follow methods similar to Girju (Girju, 2003) and Blanco (Blanco et al., 2008). We use Wordnet (University, 2010) as our lexical database. An entry of WordNet, whose gloss definition contains any of the terms in the exist-

[1] https://www.edmunds.com/recalls/

Table 2: Annotation Examples

Honda/E_1 Motor/E_1 Co./E_1 is/E_1 recalling/E_1 Acura/E_1 ILX/E_1 and/E_1 ILX/E_1 Hybrid/E_1 vehicles/E_1 because/CC_1 excessive/C_1 headlight/C_1 temperatures/C_1 pose/C_1 a/C_1 fire/C_1 risk/C_1.
Attrition/C_1 of/C_1 associates/C_1 will/CC_1 effect/CC_1 scheduled/E_1/C_2 release/E_1/C_2 of/E_1/C_2 product/E_1/C_2 causing/CC_2 high/E_2 business/E_2 impact/E_2.

ing causal list, is included in the list as a possible causal connectives. Once we have a list of words, we further expand the list by adding common phrases with contain one or more of these words. For example, the seed word *causes* is extended to include phrases like "one of the main causes of", "a leading cause of" etc. This gives us an extended connective list of 310 words/phrases. Table 3 shows a few examples of seed words and new terms added to the list. After preprocessing, we finally obtained a dataset of 8K sentences for annotation in terms of their cause, effect and causal connectives.

The Annotation Process: The above sentences are presented to three expert annotators. The experts were asked to complete the following two tasks. a) Identify whether a given sentence contains a causal event (either cause/effect) and b) Annotate each word in a sentence in terms of the four labels *cause (C)*, *effect(E)*, *causal connectives(CC)* and *None*. An illustration of the annotated dataset is depicted in Table 2.

In some of the candidate sentences, it is observed that a single sentence contains multiple cause-effect pairs, some of which are even chained together. In order to handle multiple instances of causality present in the same sentence, sentences are split into sub-sentences. e.g. *"In developing countries four-fifths of all the illnesses are caused by water-borne diseases with diarrhoea being the leading cause of childhood death"* (Hendrickx et al., 2009). This sentence has two distinct causes and their corresponding effects : *four-fifths of all the illnesses are caused by water-borne diseases* and *diarrhoea being the leading cause of childhood death.*

We have also observed a number of cases where a single sentence contains a chain of causal events where a cause event e_1 results the effect of another event e_2 which in turn causes event e_3. In such cases e_2 will be marked as both effect for e_1 and cause for e_3. For example, in *"The reactor meltdown caused a chain reaction that destroyed all the towers in the network"* (Hendrickx et al., 2009), there are two different causalities, chained

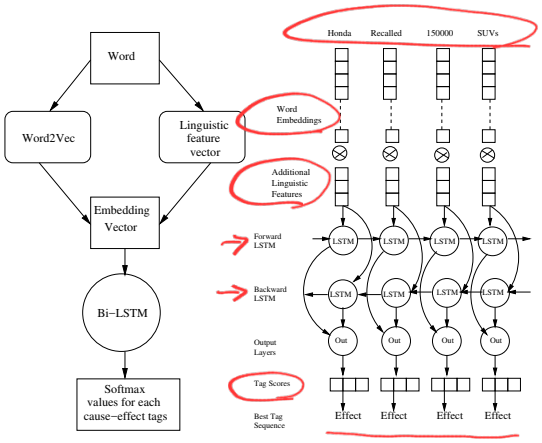


Figure 1: Overview of the bidirectional LSTM architecture for Cause-Effect relation extraction.

together: (1)*The reactor meltdown caused a chain reaction* and (2)*a chain reaction that destroyed all the towers in the network*. The effect in the first case and the cause in the second is *"A chained reaction"*. Similar example illustrated with an annotation is depicted in example (2) of Table 2. In order to extract all instances of causality present in a sentence, the sentence is divided into sub-sentences. We use openIE (Schmitz et al., 2012) to extract multiple relationships from the sentence, and then treat each relationship as a separate sentence.

Based on the given annotation scheme, each of the annotator received around 2500 sentences. Out of these, 2000 sentences are unique and rest 500 are overlapping. Using these 500 common sentences, we measure the inter annotator agreement of the annotation using the Fleiss Kappa (Fleiss and Paik, 1981) measure (κ). This is computed as $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$. The factor $1-\bar{P}_e$ gives the degree of agreement that is attainable above chance, and $\bar{P}-\bar{P}_e$ gives the degree of agreement actually achieved above chance. We have achieved the inter annotator agreement to be around 0.63. This implies that the expert annotated dataset is reliable to be used for further processing. Some more examples of annotated sentences are elaborated in the appendix A.

Table 3: Examples of seed and learnt terms from WordNet for lexical patterns

Seed	New Term	Wordnet Gloss of Term	Example
due to	corrode	cause to deteriorate due to agent	The acid corroded the metal.
	break down	collapse due to agent	Stomach juices break down proteins.
cause to	choke	become or cause to become obstructed	He choked on a fishbone.
	confuse	cause to be unable to think clearly	The sudden onslaught confused the enemy.

3.2 The linguistically informed Bi-directional LSTM model

There is a recent surge of interest in deep neural network based models that are based on continuous-space representation of the input and non-linear functions. Thus, such models are capable of modeling complex patterns in data and since they do not depend on manual engineering of features, they can be applied to solve problems in an end-to-end fashion. On the other hand, such neural network models fails to consider the latent linguistic characteristics of a text that can play an important role in extraction of the relevant information. Therefore, we have proposed a deep neural network model based on the bidirectional long-short term memory (LSTM) model (Hochreiter and Schmidhuber, 1997) (Schmidhuber et al., 2006) that along with the word embeddings, utilizes different linguistic features within a text for the automatic classification of cause-effect relations.

In identification of causal relationships from text, the surrounding context is of paramount information. While typical LSTMs allow the preceding elements to be considered as context for an element under scrutiny, we prefer to use bidirectional LSTMs (Bi-LSTM) networks (Graves et al., 2012) that are connected so that both future and past sequence context can be examined, i.e. both preceding and succeeding elements can be considered.

The overview of the proposed model is depicted in Figure 1. Corresponding to each input text, we determine the word embedding representation of each words of the text and the different linguistic feature embeddings. The input to the Bi-LSTM unit is an embedding vector (E)which is the composition of the word embedding representation (W_e) and the linguistic feature embeddings (W_l). This is represented as $\overrightarrow{E} = \overrightarrow{W_e} \bigotimes \overrightarrow{W_l}$

Generating Word Embeddings: Pre-trained GloVe word vector representations of dimension 300 have been used for this work (Pennington et al., 2014). GloVe is a relatively recent method of obtaining vector representations of words and has been proven to be effective. Along with the GloVe vector, the embedding vector of each word is appended with the vector formed from the linguistic features that has been described in the earlier section.

Generating linguistic feature embeddings: Apart from the presence of causal connectives mentioned earlier, other features added to make our model linguistically informed are relevant lexical and syntactic features Part of Speech(POS) tags (Manning et al., 2014), Universal Dependency relations (De Marneffe et al., 2006) and position in Verb/ Noun/ Prepositional Phrase structure. We have also used the semantic features as identified by Girju (Girju, 2003) - the nine Noun hierarchies (H(1) to H(9)) in WordNet namely, *entity, psychological feature, abstraction, state, event, act, group, possession, and phenomenon*. First, a single feature Primary Causal Class (PCC) is defined for a word w_i. If $w_i \in H_i$ where H_i is any of the nine WordNet hierarchies, $PCC = H_i$, else $PCC = null$. Another feature, Secondary Causal Class(SCC) is also defined. This takes value $H(i)$ if any WordNet synonym of the word belongs to $H(i)$, and is $Null$ otherwise. Further, we consider the dependency structure of the sentence, which gives us that w_i is dependent on word p_i. In addition to the five features described above for w_i, we also consider the same five features of p_i as part of w_is feature set. If w_i is not dependent on any other word in the sentence, then the parent features are the same as the word features. An example of the linguistic feature selection can be found in appendix A.

Network Architecture: We use a k-layer Bi-RNN, composed of k Bi-RNNs stacked, where the output of each such unit is the input to the next unit (Irsoy and Cardie, 2014). A two-layer stack of Bi-LSTMs is employed for the purpose of experiments. The model is trained with Adam optimizer (Kingma and Ba, 2014) and dropout layer with the dropout value of 0.5 for each Bi-RNN. The dropout layer reduces the problem of overfitting often seen in trained models by dropping

unit with connections to the neural network at random during the training process (Srivastava et al., 2014). The model is fit over runs of 2000 epochs, with batch size of 128. The loss is calculated as a function of the mean cross entropy generated. Each Bi-LSTM has 256 hidden layers and 1 final dense layer with softmax activation as output.

3.3 Causal Embeddings for Representing Similar Events

We have applied the proposed causal extraction technique over a large set of data from four different domains namely, Analyst Reports, Adverse Drug Effects, Business News and Product Recall News. We observe that a number of extracted causal events shows high degree of semantic similarity. For example, "*Engine breakdown*" and "*Engine failure*" represents the same semantic sense. Therefore, we intend to group these events into clusters. Accordingly, we device a novel algorithm to determine similar causal events. The algorithm follows the following steps: a) first identify the word embeddings of each constituent word of a causal event. The word embeddings are identified using the standard GloVe representations (Pennington et al., 2014). Apart from the word embeddings, we have also created phrase embeddings by computing a tensor product between the individual word embeddings. For example, given two causal events $C_1 = w_1, w_2 ..., w_i$ and $C_2 = w'_1, w'_2, ..w'_j$, where $w_1, w_2, ...w_k$ and $w'_1, w'_2 ...w'_k$ are the constituent word embeddings of the causal events C_1, and C_2 such that $i \neq j$, the phrase embedding $P(w_1, w_2)$ is created by computing the tensor product of each adjacent word embedding pairs. This is represented as $P(w_1, w_2) = w_1 \bigotimes w_2$. Similar word and phrase embeddings are constructed for causal event C_2. Consequently, we define A *and* B as the number of word embeddings in C_1 *and* C_2 respectively. Similarly, A' *and* B' are the number of phrase embeddings in C_1 *and* C_2 respectively. Therefore, the similarity

$$S(C_1, C_2) = \frac{(S' + S'')}{N_1 + N_2}$$

The expressions N_1 and N_2 implies $A \cup B$ and $A' \cup B'$ respectively. S' and S'' are computed as: $S' = \sum_{\forall w_i \in C_1} S_{w_i}$ and $S'' = \sum_{\forall p_i \in C_1} S_{p_i}$ Where,

$$S_{w_i} = \max_{\forall w'_j \in C_2} (Sim(w_i, w'_j))$$

$$S_{p_i} = \max_{\forall p'_j \in C_2} (Sim(p_i, p'_j))$$

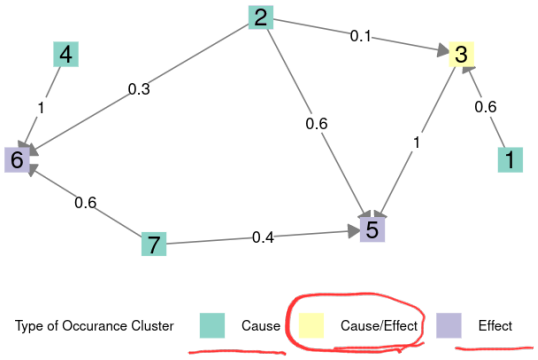
Again, p and p' are the individual phrase embeddings in sentence C_1 and C_2 respectively. $Sim(x, y)$ is the cosine similarity between the two word vector w_x and w_y. Based on the similarity score, we perform a k-means clustering to form clusters of similar causal events. We have used the Average silhouette method to identify number of clusters k. For the present work we obtained the value of k as 21. A partial network of a few representative clusters, as obtained from the vehicle Recall database, is shown in Figure 2. For each cluster, the size is given as number of phrases that constitute the cluster, and a few representative phrases of each cluster is also shown as reference. The name of the cluster is chosen from the most common noun chunks present in the cluster. The network itself is shown as a directed graph, with edges directed from Cause to Effect, as edge weights being computed as the fraction of total occurrences of the cause that lead to the effect.

Following the method each cluster can be further represented by a verb-noun pair as proposed in (Zhao et al., 2017). For noisy clusters where no such generalization is possible are left out for the time being.

4 Experiments and Results

We perform a number of different experiments to evaluate and compare the performance of our proposed system with the baseline systems. In general we classify the experiments into three different groups. Each group uses different techniques to identify causality in text. Group-1 uses rule based method, group-2 uses a CRF based classification model, group-3 uses Bi-LSTM model and group-4 uses our proposed linguistically informed Bi-LSTM model. The outputs of the experiments are evaluated in terms of the five given datasets that are explained earlier. Again, corresponding to each group, we define three different evaluation tasks. The tasks are distinguished in terms of the way each datasets are divided for training, development and testing purposes.

In Task-I, we took the five datasets separately and each dataset is divided into 80%, 10% and 10% for training, testing and development respectively. The F1 scores obtained by each system on the datasets by this model are reported in Table 4 for identified Cause, Effect and Causal Connec-



Cluster #	Representative Phrase	Size	Phrases Representative of the cluster	Occur -ance
1	Line Failure	5	a fuel-line problem; brake lines fail;	Cause
2	Fuel Tank Defect	13	potentially faulty fuel tank; defective fuel tanks pose;	Cause
3	Fuel Leak	16	possible fuel leak;fuel leakage;	Cause/ Effect
4	Fuel Pump Problem	10	problem with fuel pump; potential overheating in fuel pump;	Cause
5	Fire Risk	34	risk of fire; potential fire hazard;	Effect
6	Stalling	15	risk of staling;prevent vehicle from starting	Effect
7	Ignition switch Problem	10	ignition switch defect; ignition switch may turn off;	Cause

Figure 2: A projection of the network of cause-effect clusters

Table 4: Comparing F-scores of the Cause (C), Effect(E) and Connective (CC) extraction by the four classification models namely, Rule based (R), CRF, Bi-LSTM(BL) and Linguistically informed Bi-LSTM model (L-BL). The evaluation criteria follows Task-II technique where The models are trained and tested on five different dataset namely, Analyst Report (AR), BBC News (BBC), SemEval data (SEM), Adverse Drug Effect data(ADE) and Recall News (R).

		R	CRF	BL	L-BL
C	AR	65.92	68.02	69.10	**70.12**
	BBC	61.07	68.12	70.18	**74.63**
	SEM	68.00	71.23	81.62	**84.22**
	ADE	51.18	**69.5**	64.8	65.13
	R	76.36	74.43	75.68	**78.91**
E	AR	59.14	60.45	65.13	**66.50**
	BBC	66.34	67.03	68.91	**73.48**
	SEM	69.20	76.6	78.05	**78.86**
	ADE	58.51	**76.1**	73.56	74.05
	R	77.96	78.03	78.86	**79.16**
CC	AR	57.89	58.40	59.10	**59.84**
	BBC	61.32	64.19	69.02	**72.32**
	SEM	70.23	73.22	74.87	**75.39**
	R	66.17	70.58	72.41	**74.3**

tives.

In Task-II, we combine all the five datasets together and divide the training set, development set and test sets into 80%, 10% and 10% respectively. The division in dataset follows a five-fold manner. Therefore, the 10% testing data in fold-1 is different from the 10% testing data in fold-2 or fold-3. We compute the individual results and report the average of them.

Finally, in **Task-III**, we train the model using one dataset and test it to other four models. We conducted the experiments using the designated training portions of each dataset of BBC news, Recall News, Analyst Reports and SemEval individually to train the model and then tested all the sets on each resultant model. Of these, the best results were seen to be from the model trained on the BBC dataset.

From Table 4 we observe that in most of the

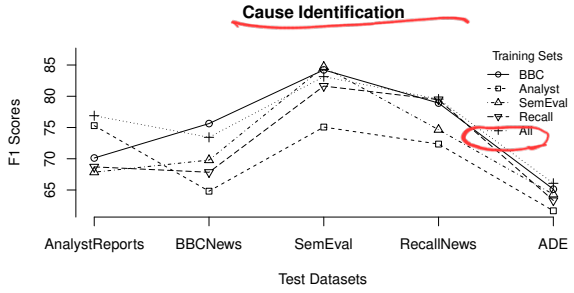


Figure 3: F1 scores for Cause Identification across different datasets for different training sets

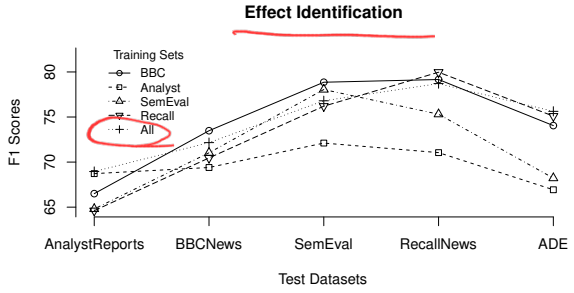


Figure 4: F1 scores for Effect Identification across different datasets for different training sets

cases Bi-directional LSTM model along with the additional layer of linguistic features significantly reduces the false negative score and achieved a high true positive score thereby achieving a high F-measure. For the project analyst report, BBC News, SEMEVAL and Recall news, we have achieved F-measures of around 66%, 73%, 79%, and 78% respectively which is best as compared to the other baseline methods. For the ADE dataset, the CRF classifier performs better than the proposed deep learning techniques, at about 73%. The inclusion of openIE as a sentence-splitter

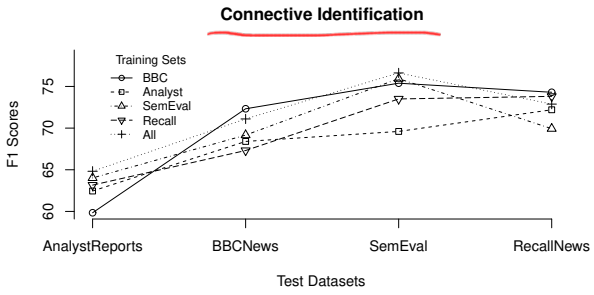


Figure 5: F1 scores for Causal Connective Identification across different datasets for different training sets

gave the most significant improvements in situations where the sentence structure was not overtly complicated, despite of the presence of multiple causal instances. Hence, the SemEval and ADE dataset results gained most from it. However, sentences from news sources often had a far more complicated structure than what OpenIE could resolve. The presence of descriptive clause along with valid cause/effect phrases made it difficult for the system to correctly identify and localize the valid phrases. In fact, the system suffered when working with such sentences, even when there was just a single instances of causality present. In the SemEval dataset, openIE usage led to identification of multiple causality in around 1/4th of the cases where multiple causality was indeed present. However, in the BBC News dataset, this amount was barely 8% of all the sentences that contained multiple instances of causation.

On an average, around 7% cases the system incorrectly predicted a cause/effect relation as valid which is actually not, whereas only 4% of the sentences were incorrectly identified as "Not an cause/effect" despite being marked as "cause/effect" by the experts. The primary reason behind this is due to fact that most of the collected texts are noisy, as a result of which the dependency parser fails to parse the texts properly and thus returning incorrect linguistic feature values. For ADE dataset, we observed that a large number of descriptions are written in languages other than English, as a result of which the classifier failed to predict correctly. Another source of error is the occurrence of incomplete sentences that restricts the classification engine to correctly label the descriptions. Apart from labeling the cause and effect events, the proposed classifier also aims to label the explicit causal connectives. Table 4 reports the results of the connective classification. We have observed that the proposed classification model is able to identify novel causal connectives that were previously not enlisted in the original causal connective list. We previously mentioned that existing schemes of having a single word represent cause and effect leads to a loss of information. Just in the SemEval dataset, just 33% of the total corpus is such that their given single-word annotation effectively captures all the information about the causal event present in the sentence. Using our proposed methodology and extending the scheme to phrases give us the complete causal information in almost 60% of the sentences that were only partially covered previously. However, we are able to somewhat quantify this observation only for the SemEval dataset, since the other datasets do not have a single-word gold standard annotation. As discussed in section 2, ambiguous causatives are a big contributor to causality being identified when it is not actually present in the sentence. Examples of some common ambiguous causal connectives, as well some of the novel connectives identified by the system (which were not present in our original list), are given in Appendix A. In addition to the above results, Figures 3, 4 and 5 show the relative performances of models trained with the individual datasets and then tested on all the test sets (Task-III).

5 Conclusion

In this paper, we present a linguistically informed deep neural network architecture for the automatic extraction of cause-effect relations from text documents. Our proposed architecture uses word level embeddings and other linguistic features to detect causal events and their effects. We evaluate the performance of the proposed model with respect to a rule based classifier and a conditional random field (CRF) based supervised classifier. We find that the bi-directional LSTM model along with an additional linguistic layer performs much better than existing baseline systems. Along with the extraction task another important contribution of this work is the development of new dataset annotated in terms of the cause-effect relations, which will be publicly shared with this paper for further research in this domain.

References

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

Quoc-Chinh Bui, Breanndán Ó Nualláin, Charles A Boucher, and Peter MA Sloot. 2010. Extracting causal relations on hiv drug resistance from literature. *BMC bioinformatics*, 11(1).

Marie De Marneffe, Bill MacCartney, Christopher Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6.

Levin B. Fleiss, J.L. and M.C. Paik. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2:212–236.

Daniela Garcia. 1997. Coatis, an nlp system to locate expressions of actions connected by causality links. *Knowledge acquisition, modeling and management*.

Roxana Girju. 2003. 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.

Roxana Girju, Dan I Moldovan, et al. 2002. Text mining for causal relations. In *FLAIRS Conference*, pages 360–364.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18. Association for Computational Linguistics.

Alex Graves et al. 2012. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384. ACM.

Ralph Grishman. 1988. Domain modeling for language analysis. Technical report, DTIC Document.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

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 *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *EMNLP*, pages 720–728.

Ashwin Ittoo and Gosse Bouma. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *International Conference on Application of Natural Language to Information Systems*, pages 52–63. Springer.

Christopher SG Khoo, Jaklin Kornfilt, Robert N Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186.

Christopher SG Khoo, Sung Hyon Myaeng, and Robert N Oddy. 2001. Using cause-effect relations in text to improve information retrieval precision. *Information processing & management*, 37(1):119–145.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Christopher Manning, Bauer Surdeanu, Mihai, Finkel John, Bethard Jenny, J. Steven, and David. McClosky. 2014. The stanford corenlp natural language processing toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning to predict from textual data. *Journal of Artificial Intelligence Research*, 45:641–684.

Ines Rehbein and Josef Ruppenhofer. 2017. Catching the common cause: extraction and annotation of causal relations and their participants. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 105–114.

Jürgen Schmidhuber, F Gers, and Douglas Eck. 2006. Learning nonregular languages: A comparison of simple recurrent networks and lstm. *Learning*, 14(9).

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

Antonio Sorgente, Giuseppe Vettigli, and Francesco Mele. 2013. Automatic extraction of cause-effect relations in natural language text. *DART@ AI* IA*, 2013:37–48.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Princeton University. 2010. Wordnet. *About WordNet*.

Fang Xuelan and Graeme Kennedy. 1992. Expressing causation in written english. *RELC Journal*, 23(1):62–80.

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 *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 335–344. ACM.

A Appendix

We use this section to elaborate on certain aspects of our work with the help of some more examples.

Table 6 shows the list of linguistic features constructed for each word of an example sentence. W1-W6 are similarly features of the original word, which are, in order,part of speech tag, universal dependency tag, parent word id, phrase structure, primary causal class and secondary causal class. Feature P is the parent word, and P1-P6 are the features of the parent word, similar to those described as W1-W6. Finally, the last column is the label associated with the word. *C* implies *Cause*, *CN* implies *Causal Connective*, *E* implies *Effect*, and *N* implies *None*.

Table 7 shows some more typical cases of causal sentences encountered and their respective annotations. As explained, the four annotation labels are *cause (C)*, *effect(E)*, *causal connectives(CC)* and *None(N)*. The second sentence contains an example of a phrase irrelevant to the actual causality that is present in the target sentence. In the current work, preciseness of the solution is dependent on it correctly disregarding the irrelevant portion and identifying causality only in the rest of the sentence. The third sentence, on the other hand, shows an example of one of the more challenging scenarios of causality identification, i.e. in the absence of any explicit causal connective. While the causality in the given sentence looks obvious to an observer, the challenge lies in the fact that there are possible grammatically and structurally similar sentences that do *not* contain causality.

Table 8 shows some common ambiguous causal connectives that identify sentences as causal even in the cases where they are not being used to identify causality. To further emphasize on their ambiguity, we show, in parallel, examples where the same connectives imply causality.

Table 5: Examples of some unusual learnt connectives

account for	**Direct payments by the patient** account for *a large proportion of funding*
derive from	*The name of Portugal* derives from **the Romano-Celtic name Portus Cale**
dictate by	*A spin label's motions* aredictated by **its local environment**
based on the fact	His *conclusion* is based on the fact **the objects contain more than 1% Arsenic**
on account of	The amount covers *expenses* on account of **his staff and transportation**
stem from	He suffers from *seizures* stemming from a **childhood injury**
punishment for	They claim the *downfall* was punishment for **the political ambitions of their leader.**
having	Having **dealt with their internal problems**, *the two companies were ripe for consolidation.*

Table 5 depicts a sample set of novel causal connectives identified by our system.

Table 6: Features of an example sentence "*Suicide is one of the leading cause of death among teens*"

Word	**W1**	**W2**	**W3**	**W4**	**W5**	**W6**	**P**	**P1-P4**	**P5**	**P6**	**Label**
Suicide	NNP	nsubj	3	B-NP	none	action	one	...	none	psychological	C
is	VBZ	cop	3	B-VP	none	none	one	...	none	psychological	N
one	CD	root	0	B-NP	none	psychological	one	...	none	psychological	CN
of	IN	case	7	B-PP	none	psychological	causes	...	none	action	CN
the	DT	det	7	B-NP	none	none	causes	...	none	action	CN
leading	VBG	amod	7	I-NP	none	action	causes	...	none	action	CN
causes	NNS	nmod	3	I-NP	action	none	one	...	none	psychological	CN
of	IN	case	9	B-PP	none	none	death	...	state	none	CN
death	NN	nmod	7	B-NP	state	none	causes	...	none	action	E
among	IN	case	11	B-PP	none	none	teens	...	none	none	N
teens	NNS	cop	9	B-NP	none	none	death	...	state	none	N

Table 7: Some typical annotation examples where causes are denoted in bold, effects are written in italic and connectives are underlined

They will *seize land owned by a British company* as part of **the President's agrarian reform program**	Example of a simple case of causality
They/*N* will/*N* seize/*E* land/*E* owned/*E* by/*E* a/*E* British/*E* company/*E* as/*CC* part/*CC* of/*CC* the/*C* President's/*C* agrarian/*C* reform/*C* program/*C*	
Gasoline is up because of **refinery issues in Texas**, which means *there will be a scramble for products in the Gulf Coast*	Example of multiple effects of single cause
Gasoline/*E1* is/*E1* up/*E1* because/*CN1* of/*CN1* refinery/*C1* issues/*C1* in/*C1* Texas/*C1* which/*CN2* means/*CN2* there/E2 will/*E2* be/*E2* a/*E2* scramble/*E2* for/*E2* products/*E2* in/*E2* the/*E2* Gulf/*E2* Coast/*E2*	
The recent falls have partly been the result of **big budget deficits**, as well as **the US's yawning current account gap**	Example of multiple causes of single effect
The/*E* recent/*E* falls/*E* have partly been the/*CN* result/*CN* of/*CN* big/*C1* budget/*C1* deficits/*C1*, as well as the/*C2* US's/*C2* yawning/*C2* current/*C2* account/*C2* gap/*C2*	
According to figures from the Ministry of Economy Trade and Industry, *the decline* was led by **a fall in demand for electronic parts for mobile phones and digital televisions**	Example of irrelevant phrase along with causal information
According/*N* to/*N* figures/*N* from/*N* the/*N* Ministry/*N* of/*N* Economy/*N* Trade/*N* and/*N* Industry/*N* the/*E* decline/*E* was/*N* led/*CC* by/*CC* a/*C* fall/*C* in/*C* demand/*C* for/*C* electronic/*C* parts/*C* for/*C* mobile/*C* phones/*C* and/*C* digital/*C* televisions/*C*	
The increase in trade has *put the country on the same level as Romania, Egypt and El Salvador*	Example with no explicit causal connective
The/*C* increase/*C* in/*C* trade/*C* has/*N* put/*E* the/*E* country/*E* on/*E* the/*E* same/*E* level/*E* as/*E* Romania/*E* Egypt/*E* and/*E* El-Salvador/*E*	

Table 8: Examples of ambiguous causatives that indicate causation only in certain context

Connective	**Example Without Causality**	**Example With Causality**
from	*The firms higher numbers* are from **improved advert sales**.	The companys sales rose to $18.6bn from last year's $12.3bn.
followed by	**The tornado** caused destruction followed by *widespread disease*.	The leader was followed by his supporters in the march.
since	*The company has cut jobs* since **demands were low**.	The company has cut 5% jobs since September 2002.