Answering Binary Causal Questions Using Role-oriented Concept Embedding

Answering Binary Causal Questions Using Role-oriented Concept Embedding

Humayun Kayesh, Md. Saiful Islam and Junhu Wang
School of Information and Communication Technology, Griffith University, Australia
{h.kayesh,saiful.islam,j.wang}@griffith.edu.au

***Abstract*—Answering binary causal questions is a challenging task, and it requires rich background knowledge to answer such questions. Extracting useful causal features from the background knowledge base and applying them effectively in a model is a crucial step to answering binary causal questions. The state-of-the-art approaches apply deep learning techniques to answer binary causal questions. In these approaches, candidate concepts are often embedded into vectors to model causal relationships among them. However, a concept may play the role of a cause in one question, but it could be an effect in another question. This aspect has not been extensively explored in existing approaches. Role-oriented causal concept embeddings are proposed in this paper to model causality between concepts. We also propose leveraging semantic concept similarity to extract causal information from concepts. Finally, we develop a deep learning framework to answer binary causal questions. Our approach yields accuracy that is comparable to or better than the benchmark approaches.**

Impact Statement—Understanding causality is crucial for automatic question-answering systems, which are useful in extracting and distributing human knowledge. An automatic question-answering system with causal knowledge can be used to check whether there is causal relationship between two concepts. Existing approaches to answer binary causal questions often answer such questions with close to 55% accuracy due to the limited usage of causal and contextual features. The deep learning framework we propose in this paper uses a role-oriented concept embedding to address such issues. Our approach achieves better accuracy by up to 3.6%, compared to the state-of-the-art benchmark approaches. The proposed approach can be used in a variety of fields, including prescriptive analysis, event prediction, and any other area where entity relationships are essential. It could also be used to improve the retrieval of causality-related inquiries in web search engines.

***Index Terms*—Causality, Causal Focus, Concept Similarity, Deep Learning, Role-Oriented Causal Embedding**

I. Introduction

BINARY questions are frequently asked for confirmation of given information. These questions can be answered by yes/no answers. Similarly, binary causal questions (BCQs) are asked to confirm whether or not a causal relationship exists between two candidate concepts. Fig. 1 depicts an example binary causal question. In the example question, "Could Covid-19 cause lung failure?," *covid-19* and *lung failure* are two candidate concepts, and the question asks whether or not there is a causal relationship between them.

Extraction and application of background knowledge are some of the key challenges in answering binary causal questions. A causal knowledge base contains causally related concepts. However, the causal relationships between concepts is directed and concepts play different roles in different contexts.

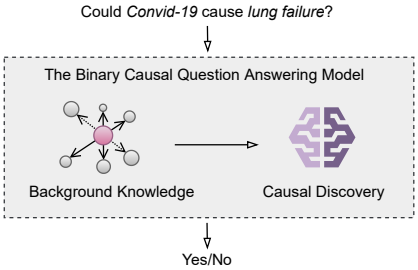


Fig. 1: Application scenario of a binary causal question answering model

In one context, a concept may play the role of a cause but in another context, the same concept may play the role of an effect. For example, in the following binary causal question, "Could accident cause death?", *accident* is a cause concept and *death* is the effect concept. However, in another binary causal question, "Could overspeeding cause accidents?", *accidents* is the effect of *overspeeding*. Therefore, a role-oriented approach to causal feature extraction is important for detecting the causal relationship between concepts in a binary causal question. Another challenge is to extract useful causal features from the causal knowledge base for each concept in a binary causal question. To extract features, a concept in a BCQ and a concept in the knowledge base must be matched. Because of the syntactic diversity of concepts in natural languages, an exact match between knowledge base concepts and input concepts is not always available.

The recent approaches to BCQs apply transfer learning-based models that are trained on automatically extracted causally-related concept pairs [1], [2]. Hassanzadeh et al. [1] proposed an approach that uses bidirectional encoder representation from transformers [3] to encode a large training dataset and then applies a top-k nearest neighbour search technique to answer binary causal questions. Later, we proposed a transfer learning-based approach [2] that was trained on a smaller dataset while keeping the performance comparable to the previous approach. However, both approaches suffer from low accuracy and precision scores. None of the approaches takes the different roles of the concepts into account. The role-oriented causal concept embedding-based approach proposed in this paper considers the context-specific roles of the candidate concepts to address this problem. We also address the challenge of causal feature extraction from a causal knowledge base using a semantic concept similarity search technique.

We assume that the candidate cause and effect concepts in a binary causal questions are already available. We aim to

answer such binary causal questions by modeling a causal relationship between the candidate cause and effect concepts. Answers to binary causal questions are binary (yes/no), so we aim to develop a deep learning framework for binary classification to answer binary causal questions. We achieve this goal by answering the following research questions.

RQ1 *How can we generate the role-oriented causal concept embeddings that can be used in causal discovery between concepts?*

RQ2 *How can we extract rich causal features from causal background knowledge base?*

RQ3 *How can we develop a deep learning framework that utilizes both contextual features and causal features to answer binary causal questions?*

In *RQ1*, we explore the options to encode the change of roles of a concept in different contexts. In *RQ2*, we study various approaches to extract causal features from the causal knowledge base. The key challenge to answer this question is to map input concepts with the causal knowledge base concepts. In the final research question, *RQ3*, we investigate the appropriate structure of a deep learning framework that effectively combines causal and effect features to model causality between concepts in binary causal questions. To be more specific, the following are our main contributions:

- we develop a novel approach to generate role-oriented causal concept embeddings;
- we propose a causal feature extraction approach using semantic concept similarity technique; and
- finally, we develop a novel deep learning framework that combines both contextual and causal features to answer binary causal questions.

The rest of the paper is organised as follows: the relevant existing approaches are discussed in Section II, the proposed approach to answer BCQs is described in Section III, the experimental settings and results are presented in Section IV and finally, the conclusions are drawn in Section V.

II. Related Work

The answering binary causal question task spans both causality detection and question-answering domains. Both of them are active research areas and we discuss notable existing approaches in these domains below.

Graph-based approaches to causality detection encode background knowledge and apply it to causal discovery [4], [5], [6], [7]. Luo et al. [4] proposed an unsupervised approach to detect causal strength between phrases. The authors collected web articles and automatically extracted causal phrases. The phrases are then split into words and a Cartesian product between the cause and effect words is performed to generate a set of causal word pairs. The authors then used these word pairs to build a directed graph. Each node in the graph corresponds to a word, and the edges represent the direction and frequency of relationships. This approach is effective for single-word input but often causal concepts contain multi-word phrases [5]. Another approach that builds causal knowledge graph of events from Wikidata [8] and Wikipedia articles was proposed by [7]. The authors used an unsupervised approach to extract causally related events from texts. However, it is challenging to extract knowledge from such causal graphs and apply it to a causality detection model.

Sharp et al. [9] proposed an approach that generates causal embeddings from free texts. The authors extracted causal tuples and trained a model for generating two context-specific embeddings for each word. Xie and Mu [10] proposed another causal embedding-based approach that generates two separate word dictionaries and their corresponding causal embeddings. The authors extracted causal phrases from raw texts and trained the embeddings generation model with them. Later, the authors improved the model using verb-mediated causal patterns [11]. However, these models only generate embeddings for single words, whereas a causal concept may contain multiple words. These works also do not address the issue of mapping input concepts to generated embeddings.

Early approaches applied rule-based techniques to detect causal relationships. Radinsky et al. [12] proposed such an approach that generates linguistic rules from the training data and the rules are then applied to detect causality. Linguistic rules are effective when input texts are formal and grammatically correct. However, they often fail when texts are informal and contain incorrect grammatical structures, such as Tweets. The above model generates rules using syntactic structures of given training data. However, learning from external knowledge bases is found to be effective in detecting causality [13]. Therefore, the recent approaches apply supervised or weakly-supervised learning in causal discovery [14], [15]. Transfer learning-based approaches are also among the popular techniques for causality detection [1], [16], [2], [17].

In the literature, the answering binary causal question task has not yet been explored extensively. One of the first approaches to answer binary causal question was proposed by Hassanzadeh et al. [1], [18]. The authors proposed a transfer learning-based approach that applies a bidirectional encoder representation from transformers model on a dataset of 17 million cause sentences to answering binary causal question. Firstly, the sentences were transformed into vectors, and then a k-nearest neighbor search was performed to find top-k sentences for a given input concept pairs. Before searching, the concept pairs were converted into two full sentences using the following templates: "X may cause Y" and "Y may cause X" where X and Y are candidate cause and effect concepts, respectively. The authors calculated the average cosine similarity of the top-k sentence with the template sentences to determine the answer to a binary causal question. The model requires two thresholds to be set based on the test data, and setting the threshold values requires knowledge of the test data. Furthermore, the model makes use of a large training dataset, which is frequently unavailable or prohibitively expensive.

In our previous work [2], we addressed these issues and proposed a transfer learning approach to answer binary causal questions. The approach fine-tunes bidirectional encoder representation from transformers-based models on a relatively smaller dataset of 100K causal pairs. The training dataset was automatically generated from news articles. However, both of the approaches achieve low accuracy and precision on evaluation datasets. For a trustworthy question-answering system, the accuracy and precision scores are crucial. As a

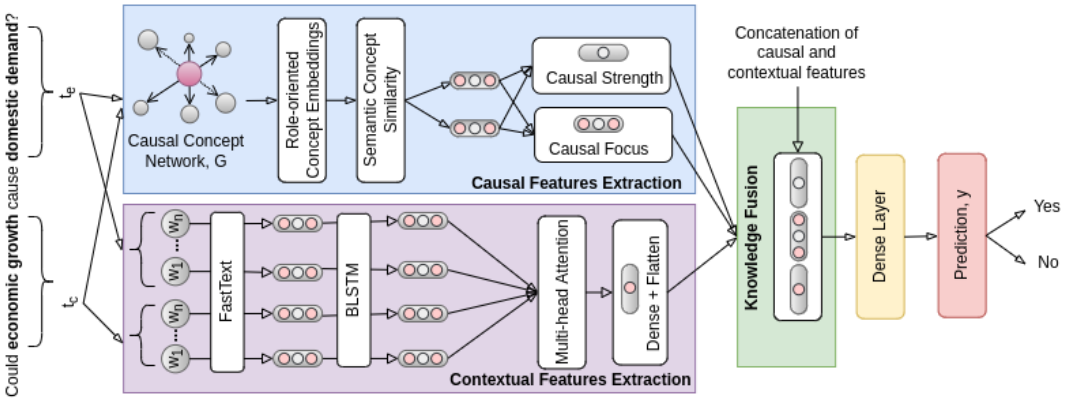


Fig. 2: The proposed deep learning framework for answering binary causal questions

result, we addressed this issue in this paper by proposing a novel deep learning framework for answering BCQs that employs a role-oriented concept embedding technique.

III. Our Approach

In this section, we describe our proposed deep learning model to answer binary causal questions. The proposed framework consists of three main components: causal feature extraction, contextual feature extraction, and knowledge fusion. Each component is described in detail below. The architecture of the proposed framework is illustrated in Fig. 2.

A. Causal Features Extraction

The causal feature extraction module extracts causal features from the background knowledge base.

1) Causal Concept Network: As the first step to causal features extraction, we build a causal concept network using a high-quality causal knowledge base. We use CauseNet [19], which is a dataset of causally related concepts published by Heindorf et al. [20], as our causal knowledge base. The authors published two versions of their dataset: CausaNet-full and CauseNet-precision. We use the latter one as it contains high precision causal concepts. The dataset consists of more than 197K concept pairs and around 80k unique concepts. From this dataset, we create a causal concept network, G. It is a directed graph where each node represents a concept and each edge represents a causal relationship between a cause and an effect concept. After preparing G, we generate embeddings for each node, v, in G. A concept may be a cause for another concept, it could also be an effect of another concept, i.e., there are two possible roles of each concept. Therefore, we need two separate embeddings to represent each role effectively.

2) Role-oriented Concept Embedding: We propose to use the alternating random walk technique [21] to generate the role-oriented embeddings of each node in G. The technique was originally proposed to generate embeddings of each node of a directed graph based on their roles. The role is determined by the inward or outward links in the directed graph. In our case, we have causal concepts as nodes and directed edges representing the orientation of the causal relationship. Fig. 3 illustrates the technique to generate role-oriented concept embeddings from G. An inward link to a node $death$ from $accident$ in G represents that $death$ is an effect of $accident$.

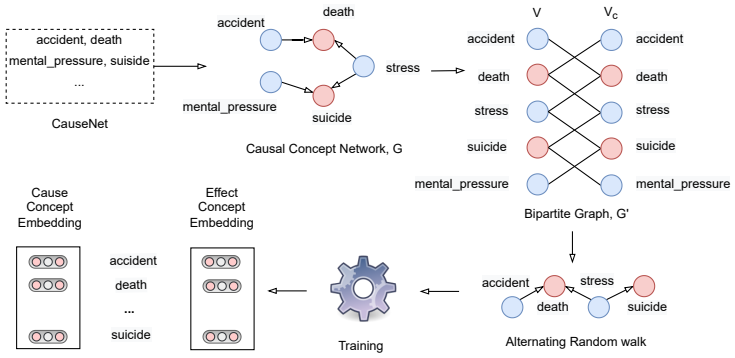


Fig. 3: Role-oriented concept embedding generation

Similarity, an outward link from a node $death$ to $stress$ represents that $death$ is a cause of $stress$.

As a first step to generate concept embedding, a bipartite undirected graph, G', is generated from G. Assume that V denotes the nodes of G. We create another set of nodes V_c that contains a duplicate of each node in G. Then we build G' between V and V_c so that there is an edge between a node from V and V_c if and only if they have an edge in G. The adjacency matrix of G' is symmetric. Then, input nodes are sampled from G and an alternating walk is performed from each input node. The walk generates sample paths with odd nodes as cause concepts and even nodes as effect concepts in order to preserve the roles of neighbourhood nodes. The alternating walk is a hybrid of two walks: the source walk and the target walk. The source walk begins with a cause node and continues with each alternating node in the path. Similarly, the effect node initiates the target walk, and each alternating node is an effect node. Finally, the model is jointly trained using the generated random walks, yielding two role-oriented embeddings for each node, v. For example, the $accident$ and $death$ nodes will have two embeddings each: cause and effect embeddings. Because $accident$ and $death$ are causally-related in G, the causal embedding of the $accident$ and the effect embedding of $death$ should be closer in the embedding space. For a thorough discussion of the alternating random walk approach, we refer the reader to [21].

3) Semantically Similar Concepts Discovery: In this step, we extract role-oriented concept embedding for an input concept pair, (t_c, t_e), from the cause and effect embeddings generated in the previous step. Here, t_c is a cause concept

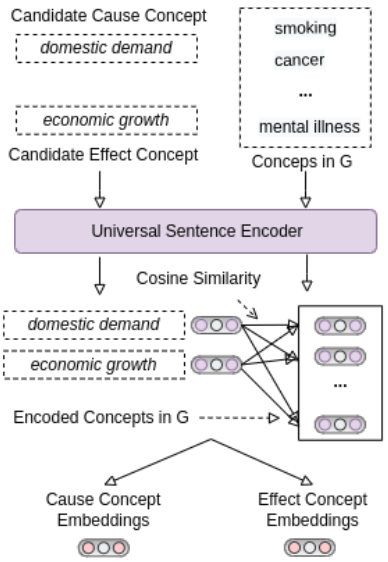


Fig. 4: Semantically similar concepts discovery

and t_e is an effect concept. Each concept may contain one or more words. The key challenge in this task is to map between an input concept and the concept nodes in G. The most obvious approach would be to apply a hit and miss scenario. If a concept matches word-to-word with a node in G, this is a hit and the corresponding embedding is returned. Otherwise, a default embedding is returned. However, due to the dynamic nature of natural languages, two semantically similar concepts can be written in many different forms. To address this challenge, we propose to use a semantic concept similarity technique to map input concepts to the those in G.

To find semantically similar concepts, first, we convert the input concepts, t_c and t_e, into vectors using a pretrained universal sentence encoder [22] model. We also encode each concept in G using the same universal sentence encoder model that is used for t_c and t_e. Then, we calculate cosine similarity between the vector of t_c and each encoded nodes in G to find the most similar concepts. Similarly, we find the semantically similar concepts in G for t_e. Then, we extract the corresponding role-oriented concept embeddings for each concept from the embeddings generated in the previous step. We denote the concept embeddings of t_c and t_e as v_c and v_e, respectively. Fig. 4 illustrates the discovery of semantically similar concepts for an example cause and effect pair.

B. Contextual Features Extraction

The inputs to this module are the same cause-effect concept pair (t_c, t_e) as the previous module. The contextual features are extracted from the concepts using a bidirectional long short-term memory [23] and a multi-head attention technique. First, we use a pretrained fastText model [24] to convert the input concepts into embeddings. Then, we apply a bidirectional long short-term memory on the cause concept embeddings and another bidirectional long short-term memory on the effect concept embeddings. The outputs of bidirectional long short-term memory layers are passed to a multi-head attention layer followed by a series of hidden layers to capture the contextual features. Each step of the proposed contextual feature extraction procedure is described in detail below.

1) Tokenization: A concept may contain one or more words. In this step, t_c and t_e are passed to a pretrained fastText tokenizer that splits each concept into tokens. The tokenizer also contains a vocabulary D which is a dictionary of tokens and their indices. The tokenizer then converts each sequence of tokens into a sequence of indices. We denote the sequence of t_c as x_c and t_e as x_e. The tokenizer uses padding and truncating techniques on the sequences to avoid variable-length inputs.

2) Embedding layer: In this step, we build embedding matrices, $R^{z \times d}$, from cause-effect sequences, using the pretrained tokenizer model, where z is the number of tokens in D and d is the embedding dimension. Each row in R corresponds to the token in the same index in D. we convert x_c and x_e into matrices by replacing each token index with its corresponding embedding in D. We denote the corresponding embedding matrices of x_c and x_e as $X_c^{l \times d}$ and $X_e^{l \times d}$, respectively, where l is the maximum length of sequence.

3) Bidirectional long short-term memory layer: The bidirectional long short-term memory layer in the proposed model is responsible for extracting context information from each concept. Following Kayesh et al. [25], we apply two separate bidirectional long short-term memory models on X_c and X_e. A bidirectional long short-term memory is consists of a forward long short-term memory and a backward long short-term memory. As shown in Eq. 1, the output of forward long short-term memory, $\overrightarrow{h_c}$, and backward long short-term memory, $\overleftarrow{h_c}$, are combined to prepare cause contextual features, h_c, where $\oplus$ denotes a concatenation operation. Similarly, the effect contextual features, h_e, are extracted as shown in Eq. 2.

$$h_c = \overrightarrow{h_c} \oplus \overleftarrow{h_c} \quad (1) \qquad h_e = \overrightarrow{h_e} \oplus \overleftarrow{h_e} \quad (2)$$

4) Attention layer: We pass the cause and effect concept features to a multi-head attention layer to extract the attention weights between the cause and effect contextual features. As shown in Eq. 3, h_c and h_e are used as the inputs to the attention layer. The output of the layer is denoted as h and f_a is a multi-head attention function proposed by Vaswani et al. [26].

$$h = f_a(h_c, h_e) \tag{3}$$

5) Dense layer: In this step, the output of the attention layer, h, is passed to a dense layer to reduces its dimension. Each neuron in this layer applies a ReLU activation function on the input as shown in Eq. 3 where W is a weight matrix b is the bias matrix.

$$h' = ReLU(h \cdot W + b) \tag{4}$$

6) Flatten layer: In this layer, we flatten the dense layer output, h', and prepare a single vector, v_m. This vector is then passed to another dense layer to generate a single value contextual feature, l, extracted from candidate cause and effect concepts. Eq. 5 shows the calculation of l where w and b_m are the weights and biases.

$$l = ReLU(v_m \cdot w + b_m) \tag{5}$$

C. Knowledge Fusion

This section describes our proposed approach for combining contextual and causal features to create a deep learning model

for answering BCQs. Creating an appropriate model structure that effectively combines two features is a challenging task. The model must be aware of the importance of the distance between concept embeddings for candidate cause and effect concepts in answering a binary causal question. The causal features extracted from G should be prioritised if a candidate cause is closer to the candidate effect in the embedding space. To this end, we propose a two-way approach for incorporating the extracted causal features into the model.

1) Causal Focus (CF): The causal features extraction step extracts v_c and v_e from t_c and t_e, respectively. To encode the interplay between cause and effect in the embedding space, we extract the causal focus features. To prepare causal focus features, r_f, we perform an element-wise multiplication between v_c and v_e as shown in Eq. 6. We find in our experiments that causal focus is an effective feature in modeling causal relationships between concepts.

$$r_f = (v_c \times v_e) \tag{6}$$

2) Causal Strength (CS): The dot product between concept embeddings, v_c and v_e, is an indicator of causal strength between the input concepts, t_c and t_e. To model this causal feature we perform a dot product of v_c and v_e and we denote the output as r_s.

After preparing the contextual feature l and the causal features r_f and r_s, we concatenate them together into a single feature vector. We allow the model to learn the weights, W_y, and biases, b_y, from the features by passing it to a dense layer that uses a Sigmoid function as the activation function. Eq. 7 shows the steps where y is the final predicted output.

$$y = \sigma((l \oplus r_f \oplus r_s) \cdot W_y + b_y) \tag{7}$$

IV. Experimental Evaluation and Analysis

This section presents our training and evaluation datasets, benchmark models, and experimental settings. Following that, we present our experimental results and compare them to the benchmark models. Finally, future research challenges to advance this field have been discussed.

A. Database Setup

We trained our model on an automatically generated database that was extracted from 1 million news articles [27]. We applied the dataset preparation technique described in Section IV-B2 and we refer this dataset as 'News Articles' in our experiments. The dataset contains 100K positive examples and an equal number of negative examples. We also trained our model on the CauseNet dataset. In this case, we automatically prepared an equal number of negative examples from the same dataset by applying the following approach.

- First, we randomly sampled the cause and effect concepts (with replacement) but swapped their positions. We used the effect concept samples as the causes and the cause concept samples as the effects in the negative examples. This technique reduces the chance of an actual causal concept pair being added to the negative examples.
- Then, we applied further filtering to remove the concept pairs that have cosine similarity scores higher than 0.2. The sampling size was twice the size of the positive concept pairs. After discarding the pairs with higher cosine similarities, we kept the size of the negative and positive examples equal. Here, we used the same role-oriented concept embeddings as described in Section III-A for the cosine similarity calculation.

We evaluated all models on the same evaluation datasets used in our previous work [2]. The evaluation datasets used in our experiments are as follows:

- **Risk Models** - This dataset was prepared from a set of decision support system models [28], [29] that represent event-event relationships as graphs. A node represents an event and an edge represents a cause-effect relationship. This dataset contains 804 cause-effect pairs.
- **CE Pairs** - This dataset is an extension of the Risk Models dataset. Seven human annotators were used to find the causes and effects of a subset of nodes in the models using a web search. The nodes and the corresponding search results were used as the cause-effect pairs. There are 302 such pairs in the dataset.
- **NATO-SFA** - This dataset was extracted from a report of Strategic Foresight Analysis (SFA) [30] published by the North Atlantic Treaty Organization in 2017. The human-generated report contains the changes to the world (causes) and their implications (effects). The dataset contains 118 such concept pairs and their labels.
- **SemEval** - This dataset is collected from SemEval 2010 sub-task 8 [31] and the size of the dataset is 1730. Each example contains a pair of labeled concepts, e.g., (collision, fire) is a pair of concepts and labeled as causal.
- **Twitter** - This dataset was prepared by Kayesh et al. [13], [15], [25] by capturing tweets related to Commonwealth Games 2018. The cause-effect pairs were extracted from Tweets using causal cue words. The dataset contains 916 pairs of causes and effects.

In each dataset, 50% of the pairs are causal and the remaining 50% are not_causal. Table I presents one causal and one not_causal pair from each dataset used in this experiment.

B. Benchmark Models

1) Existing Models in the Literature: This section describes the existing state-of-the-art models in the literature that we used as benchmark models in the experiment and compared to our proposed models.

- **PMI** - The point-wise mutual information score is the co-occurrence strength between two words. Hassanzadeh et al. [1] extended this model of calculating point-wise mutual information scores between two text spans. The model converts the candidate cause-effects concepts into two bags of phrases and then calculates the average point-wise mutual information scores between all phrase pair combinations.
- **CEA** - The cause-effect association model was proposed by Do et al. [32] that combines point-wise mutual information score with joint inverse document frequency score to calculate the causal strength between a cause-effect event pair.

TABLE I: Examples from the training and evaluation datasets

Dataset	Cause	Effect	Label
News Articles	poisonous words	homes and marriages have been destroyed	causal
	the last time jim harbaugh coached a football	game showed why he is so coveted	not causal
CauseNet	global warming	extinction	causal
	fascism	poor diet	not causal
CE Pairs	broadband access	more new businesses	causal
	increased growth	dent consumer and business confidence	not causal
NATO-SFA	consistent	Vulnerabilities	causal
	climate change	new opportunities	not causal
Risk Models	growing social tension	reduced tourism	causal
	rising regional tension	resource competition	not causal
SemEval	collision	fire	causal
	protein	researchers	not causal
Twitter	families truly suport girl-child	we can se that sky to is not the limit	causal
	i ned to be front and centre	it's al about me	not causal

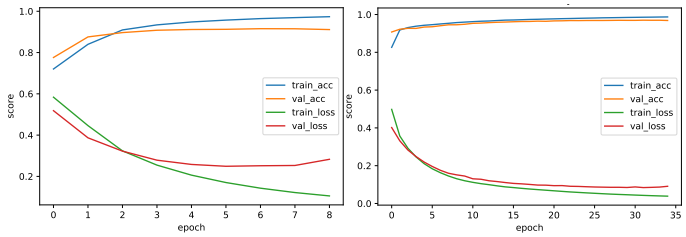
- **DCC** - This model was proposed by Hassanzadeh et al. [1]. The model indexes the list of cause-effect pairs and looks for an exact match for a candidate cause-effect pair. The model calculates the causal strength between the cause-effect pair by calculating its number of hits.
- **DCC-embed** - This model trains a customized version of word2vec model to generate phrase embeddings [1]. Then, it performs a nearest neighbour search to find the top-*k* closest cause and effect phrases. The causal relationship between the cause-and-effect pair is then calculated using the search results.
- **NLM-BERT-17M** - This neural network model was proposed by Hassanzadeh et al. [1]. The model applies a bidirectional encoder representation from transformers model to encode 17M causal sentences. Given a causal-effect concept pair (X, Y), the model converts it into two sentences: "X may cause Y" and "Y may cause X". The top-k similar sentences are then searched for each sentences and the average cosine similarity scores are calculated. To determine the causal relationship between concepts, two threshold values, one for each test dataset, are applied to the cosine similarity values. We reported the results of the above-mentioned models directly from their paper because we used the same evaluation datasets (except for the Twitter dataset) and their training dataset is not publicly accessible.
- **NLM-BERT** - This model was proposed by Hassanzadeh et al. [1]. However, we built this model ourselves and trained it on our smaller training dataset before comparing it to our proposed models. We used the same test dataset dependent thresholds as the authors for each test dataset in the evaluation.
- **NLM-BERT++** - This model is the same as NLM-BERT, but instead of selecting different thresholds for different test datasets, we used a single pair of thresholds for all test datasets. We used the automatic threshold selection technique proposed by Kayesh et al. [2]. This technique empirically discovers the optimal thresholds from the training datasets and does not depend on the test dataset.
- **CausalNet** [4]- This approach builds a causal network from training data and calculates causal scores using the network. In this experiment, we used our News Articles training data to build the causal network. In our evaluation, we consider a given candidate concept pair as causal if the causal score is greater than zero. Otherwise, we consider the pair as not_causal.

2) BERT-based Models: In our previous work [2], we proposed a transfer learning-based approach for answering binary causal question. As a part of that work, we prepared a training dataset that were extracted from news articles. We used a set of causal cue words, e.g. because, as a result of, and due to, to extract causal pairs from articles. The following example "heavy rain *causes* traffic jams" contains two causally related concepts, *heavy rain* and *traffic jams*. The full list of causal cue words used to build the dataset can be found in [2]. We extracted such concept pairs and labeled them as *causal*. To prepare the negative pairs, we collected the sentences that did not contain any causal cue words and separated them into halves. These pairs were labeled as *not_causal*.

We fine-tuned pretrained bidirectional encoder representation from transformers-based models on the training dataset mentioned-above to answer binary causal questions. Pretrained models, such as bidirectional encoder representation from transformers and RoBERTa [33], are rich in linguistic knowledge as they are trained on large datasets. Before using the training data for fine-tuning, we prepared full sentences from the concept pairs. For example, a concept pair (X, Y) is presented as "X may cause Y". Such sequences are then passed to the bidirectional encoder representation from transformers tokenizer that tokenizes the sequences and prepares them as per models accepted input format. The preprocessed sequences are then used to fine-tune the transfer learning models and the fine-tuned models were then saved for prediction and evaluation. In this paper, we use these models as the benchmark models to show the effectiveness of the proposed approach.

We implemented the following bidirectional encoder representation from transformers-based models in [2] following the technique described above to answer binary causal questions:

- **bidirectional encoder representation from transformers** [3] - bidirectional encoder representation from trans-



(a) On the News Articles dataset (b) On the CauseNet dataset

Fig. 5: Training history of our proposed model (Causal Focus + Causal Strength + Contextual Features)

formers is a transfer learning based model that was achieved state-of-the-art results in many tasks including question answering. In this experiment, we used a pre-trained bidirectional encoder representation from transformers model provided by the Huggingface [34] transformers library [35]. We used the 'bert-base-uncased' version of the pretrained model.

- **RoBERTa** [33] - RoBERTa is a variation of bidirectional encoder representation from transformers that optimizes training and hyperparameter tuning of the original model and achieves state-of-the-art results in text classification and question answering tasks. We used the 'roberta-base' version of the pretrained model from Huggingface.
- **DistilBERT** [36] - The authors of this model reduced the structural complexity and optimized the hyperparameter tuning of the original bidirectional encoder representation from transformers model. The objective of this model was to develop a lightweight version of bidirectional encoder representation from transformers without sacrificing the performance too much. We used the pretrained 'distilbert-base-uncased' version from Huggingface.
- **ALBERT** [37] - ALBERT is also a variation of bidirectional encoder representation from transformers and it was proposed focusing on scalability. The model optimizes training time and memory usage for large datasets. In our experiments, we used the 'albert-base-v2' version of the pretrained model from Huggingface.

C. Experiment Settings

We split our training datasets into training, validation, and test sets. We used 80% of the training dataset for training and the remaining 20% of the data were split into halves for validation and testing. We evaluated our model on five benchmark datasets as described in Section IV-A. We implemented our models in Python and used TensorFlow [38] and Keras [39] to construct deep learning structures. We used a pretrained fastText embedding model [40] to train our model. The fastText model was trained on the Common Crawl [41] and Wikipedia [42] data and contains 300 dimension vectors. To calculate semantic similarity, we used [43], a retrained universal sentence encoder model that returns a 512 dimension vector for the given text. We trained our model for 300 epochs while setting the patience parameter to 3. To prepare the role-oriented concept embeddings, we followed the instructions provided in Khosla et al. [21] and their code repository[1].

[1]https://git.l3s.uni-hannover.de/khosla/nerd

TABLE II: Cosine similarities between concepts

Cause	Effect	Cosine Sim	Cause	Effect	Cosine Sim
accidents	deaths	0.7690	deaths	accidents	0.3475
accidents	injury	0.7301	injury	accidents	0.7247
accidents	better_health	0.3519	better_health	accidents	-0.1230
injury	deaths	0.8054	deaths	injury	0.3521
deaths	better_health	0.2384	better_health	deaths	-0.0999
injury	better_health	0.3962	better_health	injury	-0.1395

TABLE III: Evaluation of the proposed deep learning model(s) on the test sets of the training datasets

Training Data	Models	# Pairs	Acc	Pre	Rec	F1
News Articles	CF	100K	0.5100	0.5100	**1.0000**	0.6700
	CS	100K	0.5000	0.8700	0.0000	0.0100
	Context	100K	0.8700	0.8200	0.9600	0.8800
	CF + CS	100K	0.5100	0.5200	0.2500	0.3400
	CF + Context	100K	0.9200	0.9200	0.9100	0.9200
	CF + CS + Context	100K	0.9100	0.9200	0.9100	0.9100
CauseNet	CF	197K	0.6848	0.9597	0.3930	0.5577
	CS	197K	0.7663	**1.0000**	0.5376	0.6993
	Context	197K	0.9351	0.9611	0.9084	0.9340
	CF + CS	197K	0.9483	0.9650	0.9315	0.9479
	CF + Context	197K	0.9522	0.9485	0.9574	0.9529
	CF + CS + Context	197K	**0.9698**	0.9648	**0.9758**	**0.9703**

We ran our experiments on a machine with 16 core Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz and 128GB of RAM.

The concept embeddings generation model was trained to produce two 300 dimension vectors per concept in G, a causal embedding, and an effect embedding. We set the parameters walkSize to 1, samples to 100, rho to 0.025, and joint to 0. We also set the parameter negative to 0 because there was no negative pairs in CauseNet. Rather, we trained the role-oriented causal concept embedding model on the randomly selected negative pairs. Table II displays a few examples of cosine similarities between causal and effect concept embeddings used in our experiments. Intuitively, the concepts that are causally related, such as accidents and deaths, are closer in the embeddings space, where *accidents* is the cause and *deaths* is the effect. The backward relation, where *deaths* is the case and *accidents* is the effect, has a lower cosine similarity as the concept pair is less likely to be causally related.

Because computing cosine similarity between two datasets is a computationally expensive operation, we extracted the top ten semantically similar concepts from G for each cause and effect concept in the News Articles training dataset and all evaluation datasets. In this experiment, however, we only used the top-1 concept embeddings. We did not use the semantic similarity calculation when training with the CauseNet dataset because the concepts in CauseNet already exist in G. As a result, for the CauseNet dataset, we used the exact matching technique to extract cause and effect concept embeddings.

D. Results and Discussion

1) Results on Training Dataset: We separately trained our proposed model on the News Articles and CauseNet datasets and we tested the models on the test set. The test set consists of 10% data from the corresponding training dataset. Fig. 5 shows the training history of our proposed model and Table III displays the performance of the different variations of our proposed model on the test sets of News Articles and CauseNet datasets. We denote Causal Focus, Causal Strength,

TABLE IV: Evaluation results on the CE Pairs dataset

Training Data	Models	# Pairs	Acc	Pre	Rec	F1
-	PMI	-	0.5090	0.6360	0.0440	0.0820
-	CEA	-	0.5410	0.6100	0.2250	0.3290
-	DCC	-	0.5590	0.6070	0.3380	0.4340
-	DCC-embed	-	**0.5750**	0.6110	0.4130	0.4930
-	NLM-BERT-17M	17M	0.5630	0.6470	0.2750	0.3860
News Articles	NLM-BERT	100K	0.5000	-	-	-
	NLM-BERT++	100K	0.5000	0.5000	1.0000	0.6667
	CausalNet	100K	0.5188	0.5156	0.6188	0.5625
	BERT	100K	0.5062	0.5032	0.9688	0.6624
	RoBERTa	100K	0.4750	0.4870	0.9375	0.6410
	DistilBERT	100K	0.5125	0.5064	0.9875	**0.6695**
	ALBERT	100K	0.4750	0.4868	0.9187	0.6364
	CF	100K	0.5000	0.5000	1.0000	0.6667
	CS	100K	0.5000	-	-	-
	Context	100K	0.4594	0.4764	0.8188	0.6023
	CF + CS	100K	0.5094	0.5185	0.2625	0.3485
	CF + Context	100K	0.4781	0.4868	0.8063	0.6071
	CF + CS + Context	100K	0.4438	0.4640	0.7250	0.5659
CauseNet	CF	197K	0.5000	-	-	-
	CS	197K	0.5000	-	-	-
	Context	197K	0.5375	0.5291	0.6813	0.5956
	CF + CS	197K	0.5094	1.0000	0.0188	0.0368
	CF + Context	197K	0.5344	0.5258	0.7000	0.6005
	CF + CS + Context	197K	**0.5750**	0.5723	0.5938	**0.5828**

TABLE V: Evaluation results on the NATO-SFA dataset

Training Data	Models	# Pairs	Acc	Pre	Rec	F1
-	PMI	-	0.6020	0.6760	0.3900	0.4950
-	CEA	-	0.5510	0.6500	0.2200	0.3290
-	DCC	-	0.6610	0.7020	0.5590	0.6230
-	DCC-embed	-	**0.6690**	0.7630	0.4920	0.5980
-	NLM-BERT-17M	17M	0.5590	0.5380	0.8470	0.6580
News Articles	NLM-BERT	100K	0.5000	-	-	-
	NLM-BERT++	100K	0.5000	0.5000	1.0000	**0.6667**
	CausalNet	100K	0.5339	0.5556	0.3390	0.4211
	BERT	100K	0.5000	0.5000	0.9322	0.6509
	RoBERTa	100K	0.4915	0.4956	0.9492	0.6512
	DistilBERT	100K	0.5000	0.5000	0.9661	0.6590
	ALBERT	100K	0.4661	0.4818	0.8983	0.6272
	CF	100K	0.5000	0.5000	1.0000	0.6667
	CS	100K	0.5085	1.0000	0.0169	0.0333
	Context	100K	0.5085	0.5049	0.8814	0.6420
	CF + CS	100K	0.4746	0.4681	0.3729	0.4151
	CF + Context	100K	0.5000	0.5000	0.8983	0.6424
	CF + CS + Context	100K	0.5085	0.5048	0.8983	0.6463
CauseNet	CF	197K	0.5000	-	-	-
	CS	197K	0.5085	1.0000	0.0169	0.0333
	Context	197K	0.4746	0.4851	0.8305	0.6125
	CF + CS	197K	0.5339	1.0000	0.0678	0.1270
	CF + Context	197K	0.5000	0.5000	0.7966	0.6144
	CF + CS + Context	197K	0.5508	0.5429	0.6441	0.5891

and Contextual features as CF, CS, and Context, respectively, when presenting the evaluation results. Our proposed approach performed better on the CauseNet dataset compared to the News Articles dataset when validated on the test sets of training datasets. The results in Table III suggest that in terms of accuracy and f1-score, our models perform comparatively better when trained on the CauseNet dataset. Furthermore, the CF + CS + Context model, which employs causal focus, causal strength, and contextual features, achieved the highest accuracy of 0.9698 and the highest f1-score of 0.9703.

2) Results on Evaluation Dataset: Our experimental results on evaluation datasets such as Risk Models, CE Pairs, NATO-SFA, SemEval, and Twitter are presented in the Tables IV to VIII. To evaluate the model's performance, it is trained on both News Articles and CauseNet. The model has no knowledge of the evaluation dataset. Our benchmark paper, [1], reported two types of results per model: maximum f1-score and accuracy. When comparing our models with the benchmark models proposed in [1], we compared the models with the highest accuracy values because our goal is to improve model accuracy while maintaining a balanced precision and recall.

Table IV presents the evaluation results on the CE Pairs dataset. The results suggest that the proposed model, *CF + CS + Context*, achieved the same maximum accuracy score of 0.5750 as the benchmark model DCC-embed. However, our model achieved more balanced precision, 0.5723, and recall, 0.5938 than DCC-embed. The precision of DCC-embed was better than *CF + CS + Context* but recall was below 0.50. Another benchmark, DistilBERT, achieved the highest f1-score of 0.6695 but its accuracy score of 0.5125 was comparatively low. Tables V and VI displays the evaluation results on the NATO-SFA and Risk Models datasets. Our proposed approach suffers on the NATO-SFA datasets in terms of accuracy. On the NATO-SFA dataset, the benchmark model DCC-embed achieved the highest accuracy of 0.6690 but the model's recall was below 0.5. Our best model on this dataset was *CF + CS + Context* that achieved a lower accuracy and recall than the benchmark model. However, the model's recall of 0.6441 was comparatively better and the f1-score was comparable to DCC-embed. On the Risk Models dataset, the accuracies of both benchmark models and proposed models were low. The highest accuracy score of 0.5570 on this dataset was achieved by the transfer learning-based model NLM-BERT-17M that was trained on a dataset of 17 million causal pairs. However, the proposed *CF + CS + Context* model achieved a comparable accuracy and f1-scores although the model uses a smaller training dataset of 197K causal concept pairs.

Table VII presents the evaluation results on the SemEval dataset. On this dataset, the proposed model *CF + CS* achieved the best accuracy of 0.7543 with a balanced precision and recall of 0.7397 and 0.7850, respectively. The SemEval dataset contains causal concept pairs that are similar to the CauseNet concept pairs. On this dataset, models with causal features only outperformed models with contextual features. The results on this dataset indicate that using an external knowledge base like CauseNet can significantly improve the accuracy and precision of modelling the unknown world. The benchmark model DCC-embed achieved the closest accuracy of 0.7340 compared to *CF + CS* but the recall score of this benchmark model is only 0.5790. The results on table VIII suggests that the proposed *CF + Context* model achieved the best accuracy on the Twitter dataset. Another model *CF + CS + Context* achieved a comparable accuracy of 0.5764. The benchmark model with the closest accuracy of 0.5426 was achieved by the CausalNet model but recall of this model was close to 0.5.

3) Discussion: Overall, our proposed approach achieved comparable or better accuracy over the benchmark models on all the datasets except for NATO-SFA. The proposed models that were trained on the CauseNet dataset achieved comparatively higher accuracy on the evaluation datasets (except for SemEval) compared to the proposed models trained on the news articles dataset. This finding implies that a dataset with high-quality concept pairs can improve accuracy. We

TABLE VI: Evaluation results on the Risk Models dataset

Training Data	Models	# Pairs	Acc	Pre	Rec	F1
-	PMI	-	0.5310	0.6190	0.1630	0.2580
-	CEA	-	0.5430	0.5470	0.5030	0.5240
-	DCC	-	0.5080	0.5340	0.1280	0.2060
-	DCC-embed	-	0.5200	0.5630	0.1820	0.2750
-	NLM-BERT-17M	17M	**0.5570**	0.5510	0.6140	**0.5810**
News Articles	NLM-BERT	100K	0.5000	0.5000	1.0000	0.6667
	NLM-BERT++	100K	0.5000	0.5000	1.0000	0.6667
	CausalNet	100K	0.4900	0.4938	0.7960	0.6095
	BERT	100K	0.5025	0.5013	0.9900	0.6656
	RoBERTa	100K	0.4975	0.4987	0.9701	0.6588
	DistilBERT	100K	0.4950	0.4974	0.9677	0.6571
	ALBERT	100K	0.4988	0.4994	0.9627	0.6576
	CF	100K	0.5000	0.5000	1.0000	0.6667
	CS	100K	0.5012	1.0000	0.0025	0.0050
	Context	100K	0.4677	0.4819	0.8607	0.6179
	CF + CS	100K	0.4950	0.4868	0.1841	0.2671
	CF + Context	100K	0.4851	0.4914	0.8557	0.6243
	CF + CS + Context	100K	0.4726	0.4838	0.8184	0.6081
CauseNet	CF	197K	0.5000	-	-	-
	CS	197K	0.5012	1.0000	0.0025	0.0050
	Context	197K	0.4938	0.4959	0.7438	0.5950
	CF + CS	197K	0.5025	0.7500	0.0075	0.0148
	CF + Context	197K	0.5348	0.5292	0.6318	0.5760
	CF + CS + Context	197K	**0.5323**	0.5297	0.5771	**0.5524**

TABLE VII: Evaluation results on the SemEval dataset

Training Data	Models	# Pairs	Acc	Pre	Rec	F1
-	PMI	-	0.5290	0.5300	0.5110	0.5200
-	CEA	-	0.5400	0.5420	0.5260	0.5340
-	DCC	-	0.7200	0.7930	0.5970	0.6810
-	DCC-embed	-	0.7340	0.8380	0.5790	0.6850
-	NLM-BERT-17M	17M	0.6190	0.6260	0.5930	0.6090
News Articles	NLM-BERT	100K	0.5936	0.7812	0.2601	0.3903
	NLM-BERT++	100K	0.5029	0.5014	1.0000	0.6680
	CausalNet	100K	0.5168	0.6124	0.0913	0.1590
	BERT	100K	0.5266	0.5137	0.9965	0.6779
	RoBERTa	100K	0.5006	0.5003	1.0000	0.6669
	DistilBERT	100K	0.5052	0.5026	0.9988	0.6687
	ALBERT	100K	0.5254	0.5132	0.9873	0.6754
	CF	100K	0.5000	0.5000	1.0000	0.6667
	CS	100K	0.5977	0.9721	0.2012	0.3333
	Context	100K	0.5040	0.5020	0.9977	0.6680
	CF + CS	100K	**0.7543**	0.7397	0.7850	**0.7616**
	CF + Context	100K	0.5029	0.5014	1.0000	0.6680
	CF + CS + Context	100K	0.5017	0.5009	1.0000	0.6674
CauseNet	CF	197K	0.6474	0.9923	0.2971	0.4573
	CS	197K	0.5977	0.9721	0.2012	0.3333
	Context	197K	0.4532	0.4700	0.7341	0.5731
	CF + CS	197K	0.7173	0.9372	0.4659	0.6224
	CF + Context	197K	0.5069	0.5048	0.7353	0.5986
	CF + CS + Context	197K	0.5671	0.5507	0.7283	0.6272

TABLE VIII: Evaluation results on the Twitter dataset

Training Data	Models	# Pairs	Acc	Pre	Rec	F1
	NLM-BERT-17M	17M	-	-	-	-
News Articles	NLM-BERT	100K	0.4989	-	-	-
	NLM-BERT++	100K	0.5011	0.5011	1.0000	0.6676
	CausalNet	100K	0.5426	0.5472	0.5054	0.5255
	BERT	100K	0.5044	0.5028	0.9673	0.6617
	RoBERTa	100K	0.5131	0.5073	0.9826	0.6691
	DistilBERT	100K	0.5055	0.5034	0.9630	0.6612
	ALBERT	100K	0.5142	0.5087	0.8911	0.6477
	CF	100K	0.5011	0.5011	1.0000	0.6676
	CS	100K	0.5000	1.0000	0.0022	0.0043
	Context	100K	0.4869	0.4934	0.9020	0.6379
	CF + CS	100K	0.5000	0.5022	0.2440	0.3284
	CF + Context	100K	0.4814	0.4896	0.8235	0.6141
	CF + CS + Context	100K	0.4760	0.4863	0.8105	0.6078
CauseNet	CF	197K	0.4989	0.5000	0.0022	0.0043
	CS	197K	0.5000	1.0000	0.0022	0.0043
	Context	197K	0.5284	0.5154	0.9826	**0.6762**
	CF + CS	197K	0.5000	0.6000	0.0065	0.0129
	CF + Context	197K	**0.5786**	0.5506	0.8649	**0.6729**
	CF + CS + Context	197K	**0.5764**	0.5579	0.7451	0.6381

discovered that CF and CS play an important role in detecting causality between concept pairs. They were present in all of our top-performing models on the evaluation datasets, answering our *RQ1* and *RQ2*. We also found that the models with only the causal features (CF + CF) achieved higher precision but lower recall, which resulted in lower accuracy and f1-scores. Except for the SemEval dataset, the proposed fusion technique that combines contextual and causal features outperformed the other proposed models in terms of accuracy. These findings indicate that our deep learning framework is effective at answering BCQs, which answers our *RQ3*.

E. Future Research Challenges

In this paper, our proposed approach achieved better accuracy and more balanced precision and recall compared to the benchmark models. However, there are some remaining challenges that need to be addressed to advance the field.

1) High precision, low recall: In certain cases, the benchmark models, including ours, had unbalanced precision and recall. The low recall is caused by two causality-related concepts being far apart in the embedding space. Experiments suggest that a dataset with high-quality concept pairs leads to better performance, e.g., results on the SemEval dataset as shown in Table VII. However, the development of a high-quality dataset of causally-related concepts to answer binary causal questions requires a significant effort. Because a manual approach (e.g., CauseNet) is time-consuming, an automatic extraction of high-quality causal concept pairs with performance guarantees can be investigated in the future to address this issue.

2) Concept mapping: Because our training and evaluation datasets came from different sources, it was difficult to train to map an unknown concept to one of its training concepts. We addressed this issue in part by discovering semantically similar concepts from the CauseNet for an input concept. We have applied universal sentence encoder and cosine similarity to find the most similar concepts by comparing a given concept to all the concepts in the CauseNet. This approach, however, does not always guarantee the return of causally related concepts. A better language-based modelling technique with performance guarantees can be investigated to map the training concepts to the concepts in the test datasets.

3) Quality of embeddings: We trained our role-oriented causal concept embedding model on the positive edge samples extracted from the concept network, e.g., CauseNet. When producing concept embeddings, the model automatically sampled negative examples from CauseNet. However, training the model with both true positive and negative samples may generate better embeddings, which should be investigated in the future to cover a wide range of binary causal questions.

V. Conclusion

In this paper, we proposed a novel deep learning framework to answer binary causal questions. We proposed to use the role-oriented causal embeddings of concepts and a semantic similarity technique to discover causality in text. Our proposed

approach addresses the challenge of the lack of large datasets for causality detection and demonstrates the effectiveness of role-oriented causal embeddings of concepts in improving the accuracy of the answering binary causal questions task while keeping a balanced precision and recall. Overall, our proposed approach has the potential to be useful for modelling causality in the context of binary causal questions, and it can be improved further in the future.

REFERENCES

[1] O. Hassanzadeh, D. Bhattacharjya, M. Feblowitz, K. Srinivas, M. Perrone, S. Sohrabi, and M. Katz, "Answering Binary Causal Questions Through Large-Scale Text Mining: An Evaluation Using Cause-Effect Pairs from Human Experts," in *IJCAI*, 2019.

[2] H. Kayesh, M. S. Islam, J. Wang, S. Anirban, A. S. M. Kayes, and P. A. Watters, "Answering binary causal questions: A transfer learning based approach," in *IJCNN*, pp. 1–9, IEEE, 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2018.

[4] Z. Luo, Y. Sha, K. Q. Zhu, S.-w. Hwang, and Z. Wang, "Commonsense Causal Reasoning between Short Texts," in *KR*, pp. 421–431, 2016.

[5] S. Sasaki, S. Takase, N. Inoue, N. Okazaki, and K. Inui, "Handling Multiword Expressions in Causality Estimation," *IWCS*, 2017.

[6] F. Moghimifar, G. Haffari, and M. Baktashmotlagh, "Domain adaptative causality encoder," *CoRR*, vol. abs/2011.13549, 2020.

[7] O. Hassanzadeh, "Building a knowledge graph of events and consequences using wikidata," in *Wikidata@ISWC*, vol. 2982, 2021.

[8] D. Vrandecic and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[9] R. Sharp, M. Surdeanu, P. Jansen, P. Clark, and M. Hammond, "Creating causal embeddings for question answering with minimal supervision," in *EMNLP*, pp. 138–148, ACL, 2016.

[10] Z. Xie and F. Mu, "Distributed representation of words in cause and effect spaces," in *AAAI*, pp. 7330–7337, AAAI Press, 2019.

[11] Z. Xie and F. Mu, "Boosting causal embeddings via potential verb-mediated causal patterns," in *IJCAI*, pp. 1921–1927, 2019.

[12] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality for news events prediction," in *WWW*, pp. 909–918, ACM, 2012.

[13] H. Kayesh, M. S. Islam, and J. Wang, "Event causality detection in tweets by context word extension and neural networks," in *PDCAT*, pp. 352–357, 2019.

[14] C. Hashimoto, "Weakly supervised multilingual causality extraction from wikipedia," in *EMNLP-IJCNLP*, pp. 2986–2997, ACL, 2019.

[15] H. Kayesh, M. Islam, and J. Wang, "On event causality detection in tweets," *arXiv preprint arXiv:1901.03526*, 2019.

[16] V. Khetan, R. R. Ramnani, M. Anand, S. Sengupta, and A. E. Fano, "Causal-bert : Language models for causality detection between events expressed in text," *CoRR*, vol. abs/2012.05453, 2020.

[17] V. Khetan, M. I. H. Rizvi, J. Huber, P. Bartusiak, B. Sacaleanu, and A. E. Fano, "Mimicause : Defining, identifying and predicting types of causal relationships between biomedical concepts from clinical notes," *CoRR*, vol. abs/2110.07090, 2021.

[18] O. Hassanzadeh, D. Bhattacharjya, M. Feblowitz, K. Srinivas, M. Perrone, S. Sohrabi, and M. Katz, "Causal knowledge extraction through large-scale text mining," in *AAAI*, pp. 13610–13611, 2020.

[19] "Causenet." https://causenet.org.

[20] S. Heindorf, Y. Scholten, H. Wachsmuth, A. N. Ngomo, and M. Potthast, "Causenet: Towards a causality graph extracted from the web," in *CIKM*, pp. 3023–3030, ACM, 2020.

[21] M. Khosla, J. Leonhardt, W. Nejdl, and A. Anand, "Node representation learning for directed graphs," in *ECML-PKDD*, pp. 395–411, 2019.

[22] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for english," in *EMNLP*, pp. 169–174, 2018.

[23] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[24] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[25] H. Kayesh, M. S. Islam, J. Wang, A. S. M. Kayes, and P. A. Watters, "A deep learning model for mining and detecting causally related events in tweets," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 2, 2022.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[27] D. Corney, D. Albakour, M. Martinez, and S. Moussa, "What do a million news articles look like?," in *NewsIR*, pp. 42–47, 2016.

[28] S. Sohrabi, A. V. Riabov, M. Katz, and O. Udrea, "An ai planning solution to scenario generation for enterprise risk management," in *AAAI*, 2018.

[29] S. Sohrabi, M. Katz, O. Hassanzadeh, O. Udrea, M. D. Feblowitz, and A. Riabov, "Ibm scenario planning advisor: Plan recognition as ai planning in practice," *AI Communications*, no. Preprint, pp. 1–13, 2019.

[30] NATO, "Stratigic Foresight Analysis 2017 report." https://www.act.nato.int/publications-ffao, 2017. [Online; accessed February 21, 2019].

[31] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," in *SEW*, pp. 94–99, 2009.

[32] Q. Do, Y. S. Chan, and D. Roth, "Minimally supervised event causality identification," in *EMNLP*, pp. 294–303, 2011.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[34] "Huggingface." https://github.com/huggingface/transformers.

[35] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[36] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[37] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[38] "Tensorflow." https://www.tensorflow.org/.

[39] "Keras." https://keras.io/.

[40] "Fasttext." https://fasttext.cc/docs/en/crawl-vectors.html.

[41] "Commoncrawl." https://commoncrawl.org/.

[42] "Wikipedia." https://www.wikipedia.org/.

[43] "Encoder." https://tfhub.dev/google/universal-sentence-encoder/4.

Humayun Kayesh is a final year PhD candidate at the School of ICT, Griffith University, Australia. He completed his MSc in Advanced Computer Science and IT Management at the University of Manchester in 2017. His current research interests include natural language processing, causality, social media analytics, conversational AI, and deep learning.

Md. Saiful Islam (SM'21) is a Lecturer at the School of ICT, Griffith University. He has completed his Ph.D. in Computer Science at the Swinburne University of Technology, Australia, in February 2014. He received his B.Sc. (Hons) and M.S. degree in Computer Sci. and Eng. from the University of Dhaka, Bangladesh, in 2005 and 2007, respectively. His current research interests are in the areas of database usability, AI, and big data analytics.

Junhu Wang received his Ph.D. in Computer Science from Griffith University, Australia in 2003. He is currently an associate professor at the School of ICT, Griffith University. His research interests include query processing, integrity constraint reasoning, and data analytics.