Gated Graph Neural Networks for Event Causality Identification from Social-Political News Articles

Paul Trust
University College Cork
Cork, Ireland

Rosane Minghim
University College Cork
Cork, Ireland

Kadusabe Provia
Worldquant University
Louisiana, USA

Evangelos Millos
Dalhousie University
Halifax, Canada

Abstract

The discovery of causality mentions from text is a core cognitive concept and appears in many natural language processing (NLP) applications. In this paper, we study the task of Event Causality Identification (ECI) from social-political news. The aim of the task is to detect causal relationships between event mention pairs in text. Although deep learning models have recently achieved a state-of-the-art performance on many tasks and applications in NLP, most of them still fail to capture rich semantic and syntactic structures within sentences which is key for causality classification. We present a solution for causal event detection from social-political news that captures semantic and syntactic information based on gated graph neural networks (GGNN) and contextualized language embeddings. Experimental results show that our proposed method outperforms the baseline model (BERT (Bidirectional Embeddings from Transformers) in terms of $f1-$score and accuracy.

1 Introduction

Causality is a core cognitive concept and appears in many natural language processing (NLP) tasks. We can define causality in generic terms as a semantic relationship between two arguments known as cause and effect. The occurrence of one argument (cause argument) causes the occurrence of the other (effect argument) (Feder et al., 2021; Tan et al., 2022b).

Event Causality Identification (ECI) is a task that identifies causal relationships between events from a given text (Zuo et al., 2021). To understand how documents containing causal relationships are identified, we present a sample of 5 sentences highlighting causes, effects and causal-markers leading to the rationale for classifying different documents in Figure 1 . Let us take an example of two sentences; Sentence 1: "The protests spread to 15 other towns and resulted in two death and the destruction of property" and sentence 5: "The properties including houses, banks were destroyed" as shown in Figure 1. Sentence 1 is causal and sentence 5 is non-causal. The first sentence is regraded as causal because it has the cause (in blue color) and effect (in green color) linked by a causal-marker (in red color) unlike the 5-th sentence which only has the effect.

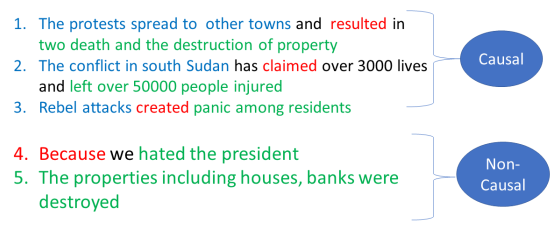

Figure 1: Examples of different text statements indicating whether they contain causal relationships or not. The causal markers are in red color, causes are in blue color and effects are in green color

In general, an expression is regarded as non-causal if any of the following conditions are satisfied; (1) the reader is unable to construct a "why" question regarding the effect, (2) the cause does not precede the effect in time, (3) the effect is equally likely to occur or not without the cause and (4) the cause and effect can be swapped without change in meaning (Tan et al., 2022b).

Event Causality Identification has been actively studied in information retrieval with deep learning as the dominant approach delivering state-of-the-art performance (Chen et al., 2015; Lai et al., 2020; Zuo et al., 2021). BERT (Devlin et al., 2019) has been utilized for automatic event causality detection on the Causal News Corpus (a dataset used in this study) (Tan et al., 2022b,a). The challenge with deep learning models is that they represent documents as a sequence of tokens either using the traditional count based methods or embedding based methods yet the task of causality detection requires understanding rich structures and reasoning within a sentence. The main contribution of this work is the use of the gated graph neural networks (GGNN) initialized with contextualized language representations on the task of causal event detection from social-political news.

2 Related Work and Background

In this section, we highlight some of the related work and background information relevant to our proposed methodology.

2.1 Document Representations

The nature in which words are represented directly influences the performance of models trained using them on downstream tasks. Traditionally, documents were represented using bag of word approaches that base on co-occurrence statistics of terms within documents (Salton et al., 1975). The key challenge with this approach is that it does not easily capture semantic relationships among words. An alternative approach to bag of words is word embeddings (Mikolov et al., 2013). Word embeddings represent words as real-valued vectors rather than counts capturing semantic and syntactic information. Word embeddings are classified into static word embeddings and contextualized word embeddings.

Static word embeddings obtain stand-alone representations of words without considering the context in which these words are used . Popular models in this category are Word2Vec models (Skip-gram and CBOW (Continuous bag of Words)) (Mikolov et al., 2013). Skip-gram uses center words to predict contextual words while CBOW uses contextual words to predict central words. GloVe (Global Vectors for Word Representation) (Pennington et al., 2014) is a log bi-linear regression model which leverages co-occurrence statistics of the corpus to represent documents. Contextual embeddings such ELMO (Peters et al., 2018) (Embeddings from Language Models) and BERT move beyond global representations like Word2Vec and assign each word a representation basing on its context hence achieving a better performance compared to static word embeddings.

2.2 Graph Neural Networks

Deep learning models especially those based on the recent transformer architecture have become dominant strategies for NLP tasks because of their impressive performance. One of the most popular transformer models is BERT (Devlin et al., 2019; Vaswani et al., 2017). BERT is a language representation model that pre-trains deep bi-directional representations from unlabeled text by jointly conditioning on both left and right contexts in all layers. It is pre-trained with two objectives: masked language modeling and next sentence prediction using the bookcorpus (800 million words) and English wikepedia (2,500 million words).

Despite the impressive performance, transformer models represent documents as a sequence of tokens which is a limitation for some NLP problems that can be naturally expressed with a graph structure. There is now a growing interest to perform deep learning on graphs using graph neural networks. Graph neural networks exploit the global features in text representations learning by aggregating information from neighbors through edges. Convolutional neural networks were first extended to handle graphs for text classification (Defferrard et al., 2016). Graph Neural Networks have since been extended to other architectures like Recurrent Neural Networks and Gated Recurrent Unit (Wu et al., 2021). In our work, we apply models graph neural networks in an application context for event causality classification from social-political news.

2.3 Event Causality Identification

The task of event causality detection from text is a semantically challenging task since it involves understanding the complex structure, relationships and dependencies within text. Traditional methods have used lexical and syntactical patterns (Hashimoto, 2019; Gao et al., 2019), co-occurrence statistics of events (Hu et al., 2017), causality markers like "due" and "because" (Hidey and McKeown, 2016) and temporal semantics of events (Ning et al., 2018). Our proposed model uses GGNN to automatically extract and induce more abstract representations.

Advanced deep learning methods based on the transformer architecture (Vaswani et al., 2017) like BERT (Bidirectional Embeddings from Transformers) (Devlin et al., 2019) have also been applied for this task (Al-Garadi et al., 2022; Nan et al., 2020). Even-though these models have achieved good per-

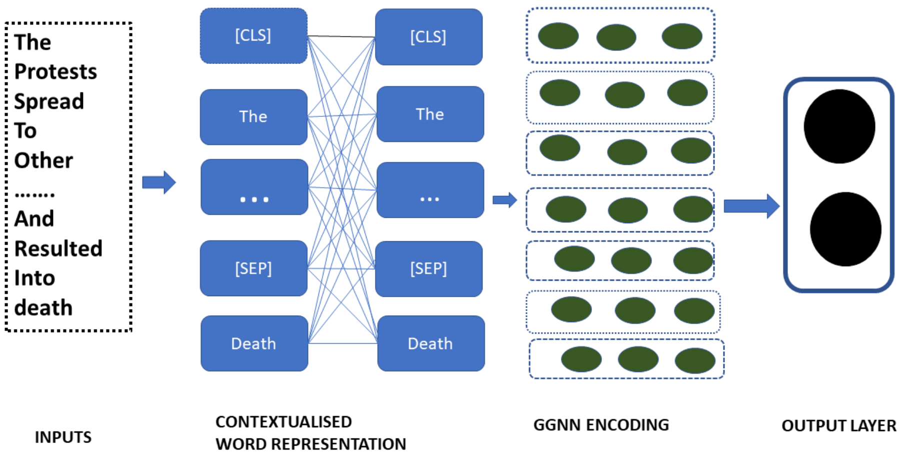

Figure 2: We first obtain contextualized embeddings of the news articles which we use to build a graph representation. A gated graph neural encoder (GGNN) and recurrent neural network decoder were used for graph neural network encoding. Finally, a fully connected neural networks was used for Event Causality Identification binary classification task

formance on event causality detection, they represent text as sequences which may not be sufficient to capture the long dependencies that are required for this event causality detection task.

Graph neural networks which extract rich structures and represent text as graph have also been explored. Graph convolutional Network (GCN) have been proposed for document level event causality detection that captures inter-sentence event mention pairs (Tran Phu and Nguyen, 2021).

Our model is different from such related work in that we use a gated graph neural network on a novel dataset; Causal News Corpus where such models have not yet as of writing the paper not explored (Tan et al., 2022b).

3 Methodology

In this section, we describe our proposed methodology for the task of Event Causality Identification from social-political news.

3.1 Document Representation

Formally, let us denote a corpus of N documents we would like to classify as $D = \{x_i, y_i\}^N$ where x_i is the i-th document with a co-responding label $y_i \in Y$ for $Y \in \{1, ..., K\}$. Each document $x_i \in D$ is represented by a sequence of words $\{w_1, ..., w_{nt}\}(w_i \in v)$ where nt is the number of words in document x_i and v is the vocabulary size.

We encode words $w_i \in x_i$ into a continuous vector representation using contextualized language representations produced by BERT (Devlin et al., 2019). Each document x_i in the corpus is represented in one token sequence which may contain a single sentence or a pair of sentences. The first token of every sequence is always a special classification token ([CLS]) and different sentences are separated by a special token ([SEP]). Documents are represented as follows [[CLS],w_1, ...w_n,[SEP],w_t,[SEP]] for an input into pre-trained BERT. We concatenate vectors of the top layers of the pre-trained BERT to obtain continuous vector representations of each word denoted as $E = \{e_i, ...e_n\}$. The embedding vectors in E are fed into a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) (Long Short Term Memory) to produce a sequence of hidden vectors $h^0 = \{h_n^0, ..., h_n^0\}$ that will be used as initialization to the graph encoder (Wu et al., 2021).

3.2 Gated Graph Neural Encoder

After representing each word in the corpus C with a corresponding word embedding, we build a graph representation of all documents in the corpus and their associated dependencies. To apply our encoder, we represent our documents as $G = (V, E)$, where V indicates a set consisting of different word embeddings for each word in the vocabulary and E indicates a set of edges (relationships) formed between documents.

We use a Gated Graph Neural network (GGNN) which is a modification of the vanilla Graph Neural Network by adding Gated Recurrent Unit filters (Chung et al., 2014). Our GGNN encoder consists of L stacked GGNN layers operating over a sequence of hidden vectors at the $i-$th layer $h^{(i)}$. The hidden vector h_i^l at the $l-$th layer is computed by averaging the hidden vectors of neighboring nodes x_i at the $(l-1)-$th layer: Gated Recurrent Unit (GRU) is used to update node embeddings by incorporating the aggregated information taking into consideration of edge type and edge direction:

$$\begin{aligned} h_i^{(0)} &= [x_i^T, 0]^T \\ a_i^{(l)} &= A_{i:}^T [h_i^{(l-1)}, ..., h_n^{(l-1)}]^T \\ h_i^{(l)} &= GRU(a_i^{(l)}, h_l^{(l-1)}) \end{aligned} \tag{1}$$

where $A \in \mathcal{R}$ is a matrix determining how nodes in the graph are communicating with each other, x_i are the initial node features, $a_i^{(l)}$ is the aggregation of information from different nodes and $h_i^{(l)}$ is the $i-$th hidden state at the $l-$th layer.

3.3 Recurrent Neural Network Decoder

The graph-level embeddings C obtained by the Graph Encoder are fed into a sequence decoder as heuristic information. In the decoding stage, an embedding layer is used to embed all the previous sequences. We used graph embedding C and sequence embedding e^t at time step t using a recurrent neural network:

$$\begin{aligned} h^t =& RNN(Concat(e^{(t)}, C), h^{(t-1)}) \\ y_t =& FC(e^{(t)}, h^{(t)}, C) \end{aligned}$$

where $h^{(t)}$ represents hidden state at time step t, $FC(.)$ represents fully connected layer and we initialize the hidden state with global graph representation C i.e $h^{(0)} = C$.

4 Experimental Results

4.1 Data

The dataset used for experiments in this paper was provided by the organizers of the shared task on Causal Event Classification organized at 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) at EMNLP 2022. The training data consists of 2925 news articles, validation set contained 323 news articles and test data consisted of 311 news articles (Tan et al., 2022b,a).

4.2 Experimental Setup

We conduct experiments with pre-trained BERT (Devlin et al., 2019) and gated graph neural networks . Experiments are done with 50 epochs, max length of 512, batch size of 50 and the learning rate was set at 0.0005. The final submissions are evaluated using $f1$-score. Transformers are implemented using hugging-face transformer library (Wolf et al., 2020) and graph neural networks were implemented using graph4nlp library (Wu et al., 2021). Our code implementation can be found on the this link (`https://github.com/TrustPaul/ggnn.git`).

4.3 Discussion

Model	**f1**	**Accuracy**
BERT (Baseline)	80.06	81.11
GGNN-W2V	81.01	75.23
GGNN-B(Ours)	**84.78**	**84.52**

Table 1: $f1-$score and accuracy on the development set of the baseline model (BERT (Bidirectional Embeddings from Transformers) and our proposed model (GGNN(Gated Graph Neural Network (Li et al., 2016; Devlin et al., 2019; Tan et al., 2022b))

Experimental results demonstrate that the performance of our proposed method (GGNN-B) compared to the baseline method that uses BERT (Devlin et al., 2019; Tan et al., 2022b) proposed by Tan et al.,(2022) as shown in Table 1. Our method improves over the baseline in terms of precision (84.78% versus 80.06%), f1 (86.19 versus 83.47%) and accuracy (84.52% versus 81.11). However fine-tuned BERT outperforms GGNN-W2V (83.47% against 76.19%) in terms of f1-score, a gated neural network of the same architecture as GGNN-B but with the graph constructed with Word2Vec embeddings.

Model	**f1**	**Accuracy**
BERT (Baseline)	78.01	77.81
GGNN-W2V	75.72	72.03
GGCN-B(Ours)	**81.67**	**80.06**

Table 2: $f1-$score and accuracy on the test set of the baseline model (BERT (Bidirectional Embeddings from Transformers) and our proposed model (GGNN(Gated Graph Neural Network (Li et al., 2016; Devlin et al., 2019; Tan et al., 2022b))

Experimental results on the test set demonstrate that our proposed method GGNN-B achieves an accuracy of 80.06% compared to an accuracy of 77.81% achieved by the baseline model (BERT). GGNN-B (our model) achieves a better f1-score compared to the baseline (82.58% against 81.12%) but BERT outperforms the same graph neural network architecture initialized with Word2Vec embeddings (Mikolov et al., 2013).

We hypothesize that the performance difference observed between our model which is based on graph neural networks and the baseline model based on only BERT is due to the superiority of graphs in representing complex structures required for understanding causal relationship against BERT that represents text as sequences. The fact that BERT outperforms Graph Neural networks when initialized with Word2Vec reinforces the role played by graph initialization of graph neural networks on performance and also demonstrates the advantages of contextualized embeddings extracted by BERT to downstream tasks over static embeddings extracted by Word2Vec.

5 Conclusion

In this work, we propose a novel deep learning approach for event causality detection from social-political news articles. Our proposed approach use gated graph neural networks and contextualized language representations which represent text documents as a graph and model complex semantic relationships ideal for causality detection. Experimental results reveal that our proposed model improves performance over the baseline comparison model (BERT) in terms of accuracy (80.06% versus 77.81%) and $f1-$score (82.58% versus 81.12%).

6 Acknowledgements

We thank Science Foundation Ireland (SFI) Center for Research Training in Advanced Networks and Future communications at University College Cork for funding this research and Irish Centre for High-End Computing (ICHEC) for providing access to computing power for running some of our experiments.

References

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, page 100217.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021. Causal inference in natural language processing: Estimation. *Prediction, Interpretation and Beyond.*

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from Wikipedia. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2988–2999, Hong Kong, China. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

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

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn Walker. 2017. Inference of fine-grained event causality from blogs and films. In *Proceedings of the Events and Stories in the News Workshop*, pages 52–58, Vancouver, Canada. Association for Computational Linguistics.

Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5405–5411, Online. Association for Computational Linguistics.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. 2016. Gated graph sequence neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Tommaso Caselli, Onur Uca, Farhana Ferdousi Liza, and Nelleke Oostdijk. 2022a. Event causality identification with causal news corpus - shared task 3, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022b. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lingfei Wu, Yu Chen, Heng Ji, and Yunyao Li. 2021. Deep learning on graphs for natural language processing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 11–14, Online. Association for Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.