Event Causality Identification via Generation of Important Context Words

Hieu Man[1], **Minh Van Nguyen**[2], **and Thien Huu Nguyen**[2]
[1] VinAI Research, Vietnam
[2] Dept. of Computer and Information Science, University of Oregon, Eugene, OR, USA
v.hieumdt@vinai.io,
{minhnv,thien}@cs.uoregon.edu

Abstract

An important problem of Information Extraction involves Event Causality Identification (ECI) that seeks to identify causal relation between pairs of event mentions. Prior models for ECI have mainly solved the problem using the classification framework that does not explore prediction/generation of important context words from input sentences for causal recognition. In this work, we consider the words along the dependency path between the two event mentions in the dependency tree as the important context words for ECI. We introduce dependency path generation as a complementary task for ECI, which can be solved jointly with causal label prediction to improve the performance. To facilitate the multi-task learning, we cast ECI into a generation problem that aims to generate both causal relation and dependency path words from input sentence. In addition, we propose to use the REINFORCE algorithm to train our generative model where novel reward functions are designed to capture both causal prediction accuracy and generation quality. The experiments on two benchmark datasets demonstrate state-of-the-art performance of the proposed model for ECI.

1 Introduction

In Information Extraction (IE), Event Causality Identification (ECI) aims to predict causal relation between a pair of events mentioned in text. For instance, in the sentence "*Massive fires cause major damages in the downtown area.*", an ECI system needs to realize the causal relation between the two events triggered by "*fires*" and "*damages*" (called event mentions), i.e., "*fires*" $\xrightarrow{\text{cause}}$ "*damages*". ECI is an important problem with many applications in NLP (Hashimoto, 2019; Berant et al., 2014).

Compared to the feature-based methods (Do et al., 2011; Ning et al., 2018), recent deep learning models have demonstrated their state-of-the-art performance for ECI (Kadowaki et al., 2019; Liu et al., 2020; Zuo et al., 2021). As such, prior work has mainly treated ECI as a classification problem where the only output from the models is a label to indicate causal or non-causal relation between input events. A major issue with this classification formulation is that current ECI models do not output important contexts for causal prediction of two event mentions. In this work, important contexts refer to the words in the input sentence that are critical to reveal the causal relation between two given event mentions (e.g., the words "*caused*" and "*by*" in our example). This limitation of current ECI models is undesirable as we expect that including important context words as a part of the outputs for ECI models can improve the training signals for the models. In particular, motivated by relation exaction models in IE (Zhang et al., 2018), we use the words along the dependency path between the two event mentions in the dependency tree to represent important context words for ECI. Our intuition is that dependency path generation is a related/complementary task for causal label prediction in ECI, and training a model to jointly generate causal labels and dependency path words (i.e., multi-task learning) can boost the performance.

A potential challenge with this idea involves the varying number of dependency path words where the generation of a context word or causal label might need to condition on previously generated ones (e.g., dependencies at the output level). As the result, such dependencies make it difficult to extend existing classification-based ECI models to perform multi-task learning with important context prediction. To address this issue, we propose to solve ECI via a new generative formulation: given a pair of event mentions in an input sentence, our ECI model aims to simultaneously generate causal label and the dependency path words between the two event mentions. In our model, causal label and dependency path words are combined into a single output sequence that will be generated by a

Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pages 323 - 330
July 14-15, 2022 ©2022 Association for Computational Linguistics

generative model from the input sentences in an autoregressive fashion, thus facilitating the encoding of dependencies between output words in our multi-task learning idea. Finally, to solve the resulting sequence-to-sequence problem for ECI, we leverage the generative pre-trained language model T5 (Raffel et al., 2020). To our knowledge, this is the first work to use generative models to solve ECI. The generation of dependency paths for relation extraction problems is also novel in IE.

Following prior work that reformulates NLP tasks into generative problems (Paolini et al., 2021; Zhang et al., 2021), we can train the generative model for ECI by maximizing the likelihood of the golden output sequences. However, this approach suffers from a potential mismatch between the used optimization objective (i.e., the likelihood) and the targeted performance measure (e.g., the accuracy for event causal prediction). In addition, as the words along the dependency paths might outnumber the causal label in the output sequence, likelihood maximization training will downgrade the importance of causal labels as a training signal in our multi-task learning framework for ECI. To this end, we propose to train our generative model for ECI using the policy-gradient method REINFORCE (Williams, 1992) that allows us to directly treat the targeted performance measure as the reward to train the generative model. Our training reward will contain separate terms for the accuracy of the predicted causal labels and the similarity of the generated and golden output sequences to allow an emphasis on the ECI performance for training. We also present a new auxiliary reward that encourages the similarity between predicted and input sentences with respect to the causal prediction ability to enrich the training signals. Finally, we conduct experiments on two benchmark datasets, demonstrating advantages of the proposed model with state-of-the-art performance for ECI.

2 Model

Given a sentence W and two event mentions e_s and e_t in W, ECI aims to predict whether e_s and e_t are involved in a causal relation in W. In this work, we depart from the traditional classification formulation (Tran and Nguyen, 2021) to a generative approach for ECI. Our generative model follows the sequence-to-sequence setting where the input sequence should capture the input sentence W along with the two event mentions e_s and e_t. In contrast, the output sentence will include the causal label and the dependency path between e_s and e_t in the dependency tree of W to achieve multi-task learning with important context word generation. To this end, the input I for our generative ECI model is obtained by combining W and a prompt $P(e_s, e_t)$ to specify the two input event mentions and the goal of ECI, i.e., $I = W : P(e_s, e_t)$. In this work, we use a simple template for $P(e_s, e_t)$ in the form of "*Is there a causal relation between e_s and e_t?*". As such, the output sequence O is then formed using the concatenation: $O = l, D(e_s, e_t)$ (called golden output). Here, l is either "*Yes*" or "*No*" to indicate the existence of a causal relation between e_s and e_t (i.e., causal label) while $D(e_s, e_t)$ represents the dependency path between e_s and e_t in W. In our example, the input and output sequences are:

I: Massive fires cause major damages in the downtown area: Is there a causal relation between fires and damages?

O: Yes, fires cause damages

Given the transformed input-output pair (I, O) for every example in the training data of ECI, we adopt the pre-trained tranformer-based language model T5 (Raffel et al., 2020) to solve the resulting sequence-to-sequence problem. In particular, we train T5 on the transformed input-output pairs (I, O) from ECI training data. At inference time, given an input sentence and two event mentions, we use the trained T5 model to generate the output sequence (with greedy decoding) from which the causal label can be extracted from the first token (i.e., l in O) to serve as the prediction.

Training: As presented in the introduction, to employ label accuracy as the direct training signal, we propose to leverage the REINFORCE algorithm (Williams, 1992) to train our T5 model for ECI where label accuracy will be used to form the reward function. In addition, the flexibility of REINFORCE allows us to include the similarity between the predicted output sequence, denoted by C, from T5 and the golden output O and input I as terms in the reward function to train our generative model. As such, we propose the following information for the reward function $R(C)$ for REINFORCE:

- **Performance-based Reward** $R^{per}(C)$: We compute this reward based on the accuracy of the causal label p in the generated sequence C (i.e., the first token of either "*Yes*" or "*No*"). In particular, $R^{per}(C) = 1$ if p is consistent with the provided relation between e_s and e_t in W, and 0 otherwise.

- **Output-based Reward** $R^{out}(C)$: This reward aims to encourage the similarity between the generated sequence C and the golden output sequence O to train the generative model T5. As such, we employ the ROUGE-2 measure (Lin, 2004) between C and O for this reward term: $R^{gold}(C) = \text{ROGUE-2}(C, O)$[1].

• **Input-based Reward** $R^{in}(C)$: Our goal is to generate the dependency path between e_s and e_t for multi-task learning for ECI. Given that the dependency path is expected to contain important contexts in W to reveal the causal relation and the input I is customized for the causal prediction purpose, we argue that the input and output sequences I and O should have similar meanings. Based on that intuition, we introduce a novel reward term $R^{in}(C)$ to promote the similarity between the generated sequence C from T5 and the input sequence I. In particular, we first send C and I (prepended with the special tokens </s>) to the encoder of T5. The vectors for </s> in the last transformer layer for C and I are then used for their representation vectors $V(C)$ and $V(I)$ respectively. Finally, the reward $R^{in}(C)$ is computed via the representation similarity, i.e., $R^{in}(C) = cosine(V(C), V(I))$.

Consequently, the overall reward function $R(C)$ to train our T5 model for ECI is: $R(C) = \alpha_{per}R^{per}(C) + \alpha_{out}R^{out}(C) + \alpha_{in}R^{in}(C)$ (α_{per}, α_{out}, and α_{in} are trade-off parameters). In this way, we can explicitly make sure that label accuracy (i.e., our main performance goal) is well represented and not dominated by the generation rewards in the training. Let $P(C|I)$ be the distribution over generated sequences that T5 induces. In our model, REINFORCE trains T5 by minimizing the negative expected reward $R(C)$ over the possible choices of C from T5: $\mathcal{L} = -\mathbb{E}_{C' \sim P(C'|I)}[R(C')]$. Using policy gradient and one roll-out sample with the generated sequence C, the gradient of $\mathcal{L}$ can be estimated for training via: $\nabla\mathcal{L} = -(R(C) - b)\nabla \log P(C|I)$ where b is a baseline to reduce variance. Here, we obtain the baseline b via: $b = \frac{1}{|B|}\sum_{q=1}^{|B|} R(C^q)$, where $|B|$ is the mini-batch size and C^q is the generated sequence for the q-th sample.

Finally, before REINFORCE training, we first bootstrap T5 by training it over the transformed pairs (I, O) with maximum likelihood objective. This helps constrain the large action space with text generation to improve the learning for REINFORCE (Ranzato et al., 2016; Paulus et al., 2018).

3 Experiments

Datasets and Hyperparameters: We evaluate our proposed generative model, called **GenECI**, on two benchmark English datasets for ECI, i.e., EventStoryLine and Causal-TimeBank. Proposed by (Caselli and Vossen, 2017), EventStoryLine (i.e., version 0.9) involves 258 documents, 22 topics, 4316 sentences, 5334 event mentions, and 1770 of 7805 event mention pairs with causal relation in a sentence. Following the same data split in previous work (Tran and Nguyen, 2021; Zuo et al., 2021), we utilize the last two topics in EventStoryLine for the development data while the remaining 20 topics are used for 5-fold cross-validation evaluation. For Causal-TimeBank (Mirza, 2014a), there are 184 documents, 6813 event mentions, and 318 of 7608 event mention pairs annotated with causal relation. Using the same setting and data split as previous work (Liu et al., 2020; Zuo et al., 2021), we perform 10-fold cross-validation evaluation.

We tune the hyperparameters for GenECI on the development data of EventStoryLine; the chosen parameters are employed to train the models for both EventStoryLine and Causal-TimeBank. The selected hyperparameters from our tuning process involve: $5e$-5 for the learning rate with the Adam optimizer; 32 for the mini-batch size; and 1.0, 0.5 and 0.1 for the trade-off-parameters α_{per}, α_{out} and α_{in} (respectively) in the overall reward function $R(C)$. Finally, we use the base version of T5 (Raffel et al., 2020) for the generative model in this work.

Comparison: We compare our model with the state-of-the-art (SOTA) models for ECI. For EventStoryLine, we consider the following baselines: (1) **LSTM** (Gao et al., 2019) adopted from (Cheng and Miyao, 2017); (2) **Seq** (Gao et al., 2019) adopted from (Choubey and Huang, 2017) for ECI; and (3) **LR+** and **LIP** (Gao et al., 2019): document structure-based models for ECI. For Causal-TimeBank, we evaluate **RB**: a rule-based system in (Mirza, 2014b), and **ML**: a feature-based model for ECI in (Mirza, 2014a). For both datasets, we also compare with the following BERT-based models for ECI: (i) **BERT**: a BERT-based baseline in (Zuo et al., 2021); (ii) **KnowDis** (Zuo et al., 2020): a model with distant supervision; (iii) **Know** (Liu et al., 2020): a model with ConceptNet; (iv) **RichGCN** (Tran and Nguyen, 2021): a graph convolutional network with rich information, and (v) **LearnDA** (Zuo et al., 2021): a data augmentation

[1] We have tried BLUE, METEROR, and other variants of ROUGE; however, ROUGE-2 leads to the best performance.

method. **RichGCN** has the best reported performance on EventStoryLine while **LearnDA** is the current SOTA model for Causal-TimeBank. Finally, we also report the performance of **T5 Classify** that is similar to the classification-based model **BERT** (Zuo et al., 2021), but replaces the BERT encoder with the encoder from T5.

Model	EventStoryLine			Causal-TimeBank		
	P	R	F1	P	R	F1
LSTM	34.0	41.5	37.4	-	-	-
Seq	32.7	44.9	37.8	-	-	-
LR+	37.0	45.2	40.7	-	-	-
LIP	37.4	55.8	44.7	-	-	-
RB	-	-	-	36.8	12.3	18.4
ML	-	-	-	67.3	22.6	33.9
BERT	36.1	56.0	43.9	38.5	43.9	41.0
KnowDis	39.7	66.5	49.7	42.3	60.5	49.8
Know	41.9	62.5	50.1	36.6	55.6	44.1
RichGCN	49.2	63.0	55.2	39.7	56.5	46.7
LearnDA	42.2	69.8	52.6	41.9	68.0	51.9
T5 Classify	39.1	69.5	47.7	39.1	67.7	48.3
GenECI (ours)	59.5	57.1	**58.8**	60.1	53.3	**56.5**

Table 1: Model performance on two datasets.

Table 1 presents the performance of the models on two datasets. The most important observation is that GenECI significantly outperforms ($p < 0.01$) the baseline models with substantial gaps on both datasets (e.g., 3.6% better than the second best model RichGCN on EventStoryLine using F1 score). Compared to "*T5 Classify*" that uses the same encoder as GenECI, it is clear that the generation-based approach with T5 is more beneficial for ECI than the classification-based method. In addition, we note that the baseline models for ECI often need external knowledge resources (e.g., ConceptNet) or additional training data (e.g., via data augmentation) to improve the performance. Our generative model does not require such resources to achieve the best performance.

Line	Model	P	R	F1
1	**GenECI** (full)	59.5	57.1	**58.8**
2	GenECI - $R^{per}(C)$	59.8	49.3	53.4
3	GenECI - $R^{out}(C)$	50.3	59.8	56.9
4	GenECI - $R^{in}(C)$	49.5	60.9	56.1
5	GenECI - ML pre-training	49.1	62.4	57.3
6	GenECI - dep path	57.0	53.9	55.4
7	Only ML training	60.0	53.5	55.7
8	Only ML training with no dep path	56.5	45.6	50.1

Table 2: Ablation study.

Input Sentence	GenECI	ML Train
Iranian rescue workers handed out blankets, food and water Monday to survivors of a powerful **earthquake** *on a Gulf island that killed 10 people and forced villagers to* **spend** *the night in tents.*	Yes, earthquake killed forced spend	No, earthquake survivors handed forced spend
Power was **restored** *to the afflicted villages on the Gulf island of Qeshm after a blackout caused by the* **quake**, *which struck on Sunday with a force of about 6.0 on the Richter scale.*	No, restored blackout caused quake	Yes, restored caused quake

Table 3: Examples with successful generation of causal labels from GenECI and incorrect generation of causal labels from ML Training. Event mentions are highlighted. ML Training generates incorrect dependency paths that include irrelevant/noisy words (e.g., "*survivors*" and "*handed*" in the first example) or miss important context words (e.g., "*blackout*" in the second example). Such missing or irrelevant information suggests inability to encode important context for successful causal label prediction.

Ablation Study: This section studies the contribution of each designed component for GenECI. In particular, the major components in GenECI involve the dependency path generation, the REINFORCE training with different reward terms, and the maximum likelihood (ML) pre-training. Table 2 shows the performance of the ablated models on the test set of EventStoryLine when the components are eliminated from GenECI. As can be seen from lines 2, 3, 4, and 5, the proposed reward functions $R^{per}(C)$, $R^{out}(C)$, $R^{in}(C)$ and the ML pre-training are all important to produce best performance for GenECI. In line 6, we exclude the dependency paths from the output sequences O (i.e., O only contains the causal label), which essentially amounts to not using multi-task learning with dependency path generation for GenECI. This also leads to the exclusion of the reward terms $R^{out}(C)$ and $R^{in}(C)$ from $R(C)$. It is clear from the table that the performance of GenECI suffers significantly due to the dependency path removal, verifying the effectiveness of multi-task learning with dependency paths for ECI. Next, in lines 7 and 8, we present the performance of T5 when it is only trained with the ML objective. As the performance of ML training is substantially worse, it suggests that REINFORCE training with the designed rewards is more effective for generative ECI.

Analysis: To better understand the operation of GenECI, we analyze the examples in EventStoryLine that are successfully predicted by GenECI, but cannot be recognized correctly by the ML training model (i.e., only training T5 with maximum likelihood objective). Our main finding from the analysis is that GenECI can generate correct dependency paths between two given event mentions that demonstrates the ability to learn necessary context for successful prediction. In contrast, ML training tends to produce incorrect dependency paths (i.e., including irrelevant words or missing important words), thus showing limited representation learning ability and leading to causal prediction failure. Table 3 presents two examples to demonstrate the effectiveness of GenECI and reveal issues for ML Training.

4 Related Work

In the early methods, ECI has been mostly approached by feature-based models (Beamer and Girju, 2009; Do et al., 2011; Riaz and Girju, 2014; Hidey and McKeown, 2016; Ning et al., 2018; Hashimoto, 2019; Gao et al., 2019). Recently, ECI has been further solved by deep learning models (Gao et al., 2019) where external knowledge and additional training data are leveraged to improve the performance (Liu et al., 2020; Zuo et al., 2020, 2021; Tran and Nguyen, 2021). We are different from such prior work as we are the first to model ECI via a generative model.

Using generative models for traditional classification-based problems has also been explored recently, e.g., for named entity recognition (Athiwaratkun et al., 2020; Yan et al., 2021), sentiment analysis (Zhang et al., 2021), and event extraction (Lu et al., 2021). However, none of such prior work considers generative models for ECI. Finally, we also note related work on extracting other types of relations between event triggers, including temporal relation (Ning et al., 2017; Leeuwenberg and Moens, 2017; Ning et al., 2018b; Tran Phu et al., 2021), subevent relation (Glavaš et al., 2014; Araki et al., 2014; Aldawsari and Finlayson, 2019; Man et al., 2022), and coreference relation (Nguyen et al., 2016; Choubey and Huang, 2018; Huang et al., 2019; Choubey et al., 2020; Phung et al., 2021; Minh Tran et al., 2021).

5 Conclusion

We introduce a novel model for ECI that solves the problem via a generation framework with the T5 model. Our model explores multi-task learning that jointly generates the dependency paths between two event mentions for ECI. We also introduce a training procedure based on REINFORCE and novel reward functions, which leads to the SOTA performance for ECI. In the future, we plan to extend the model to other relation extraction tasks.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112 and the NSF grant CNS-1747798 to the IUCRC Center for Big Learning. This research is also based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ARO, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

References

Mohammed Aldawsari and Mark Finlayson. 2019. Detecting subevents using discourse and narrative features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jun Araki, Zhengzhong Liu, Eduard Hovy, and Teruko Mitamura. 2014. Detecting subevent structure for event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020. Augmented natural language for generative sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *CICLing*.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter

Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1796–1802, Copenhagen, Denmark. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2018. Improving event coreference resolution by modeling correlations between event coreference chains and document topic structures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

Goran Glavaš, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

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.

Yin Jou Huang, Jing Lu, Sadao Kurohashi, and Vincent Ng. 2019. Improving event coreference resolution by learning argument compatibility from unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. 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 5816–5822, Hong Kong, China. Association for Computational Linguistics.

Artuur Leeuwenberg and Marie-Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3608–3614. International Joint Conferences on Artificial Intelligence Organization.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*.

Hieu Man, Nghia Trung Ngo, Linh Van Ngo, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the Conference on the Advancement of Artificial Intelligence (AAAI)*.

Hieu Minh Tran, Duy Phung, and Thien Huu Nguyen. 2021. Exploiting document structures and cluster consistencies for event coreference resolution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4840–4850, Online. Association for Computational Linguistics.

Paramita Mirza. 2014a. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17, Baltimore, Maryland, USA. Association for Computational Linguistics.

Paramita Mirza. 2014b. Fbk-hlt-time: a complete italian temporal processing system for eventi-evalita 2014. In *EVALITA*.

Thien Huu Nguyen, , Adam Meyers, and Ralph Grishman. 2016. New york university 2016 system for kbp event nugget: A deep learning approach. In *Proceedings of the Text Analysis Conference (TAC)*.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018b. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero dos Santos Nogueira, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Duy Phung, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2021. Hierarchical graph convolutional networks for jointly resolving cross-document coreference of entity and event mentions. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 32–41, Mexico City, Mexico. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mehwish Riaz and Roxana Girju. 2014. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *SIGDIAL*.

Minh Phu Tran 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.

Minh Tran Phu, Minh Van Nguyen, and Thien Huu Nguyen. 2021. Fine-grained temporal relation extraction with ordered-neuron LSTM and graph convolutional networks. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 35–45, Online. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Kluwer Academic*.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3558–3571, Online. Association for Computational Linguistics.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1544–1550, Barcelona, Spain (Online). International Committee on Computational Linguistics.