KnowDis: Knowledge Enhanced Data Augmentation for Event Causality Detection via Distant Supervision

Xinyu Zuo[1,2], **Yubo Chen**[1,2], **Kang Liu**[1,2], **Jun Zhao**[1,2]
[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
[2]School of Artificial Intelligence,
University of Chinese Academy of Sciences, Beijing, 100049, China
{xinyu.zuo,yubo.chen,kliu,jzhao}@nlpr.ia.ac.cn

Abstract

Modern models of event causality detection (ECD) are mainly based on supervised learning from small hand-labeled corpora. However, hand-labeled training data is expensive to produce, low coverage of causal expressions and limited in size, which makes supervised methods hard to detect causal relations between events. To solve this data lacking problem, we investigate a data augmentation framework for ECD, dubbed as **Know**ledge Enhanced **Dis**tant Data Augmentation (**KnowDis**). Experimental results on two benchmark datasets *EventStoryLine corpus* and *Causal-TimeBank* show that 1) KnowDis can augment available training data assisted with the lexical and causal commonsense knowledge for ECD via distant supervision, and 2) our method outperforms previous methods by a large margin assisted with automatically labeled training data.

1 Introduction

Event causality detection (ECD) aims to identify causal relations between events from texts, which may provide crucial clues for many NLP tasks, such as information extraction, logical reasoning, question answering, and others (Girju, 2003; Oh et al., 2013; Oh et al., 2017). For example, the causal relation that Kimani Gray was *killed* because of a police *attack* is needed to be detected in the following sentence: *"Kimani Gray, a young man who likes football, was **killed** in a police **attack** shortly after a tight match."*

This task is usually modeled as a classification problem, i.e. determining whether there is a causal relation between two events in a sentence. To this end, most existing methods adopt a supervised learning paradigm (Mirza and Tonelli, 2016; Riaz and Girju, 2014; Hashimoto et al., 2014; Hu and Walker, 2017; Gao et al., 2019; Zuo et al., 2020). Although these methods have achieved good performance, they usually need large-scale annotated training data. However, existing event causality detection datasets are relatively small. For example, the EventStoryLine Corpus (Caselli and Vossen, 2017) only contains 258 documents, 4316 sentences, and 1770 causal event pairs. These small datasets are in low coverage of causal expressions and obstacle NLP applications deployed on large-scale data. Recent improvements of distant supervision have been proven to be effective to label training data for some tasks, such as relation extraction (Mintz et al., 2009), event detection (Chen et al., 2017), and so on. Therefore, we investigate a distant data augmentation framework for solving the data lacking problem on the ECD task, dubbed as **Know**ledge Enhanced **Dis**tant Data Augmentation (**KnowDis**), to automatically label available data.

We argue that a sentence contains an event pair with a high probability of causality and expresses its causal semantic can be labeled as training data for the ECD task. To automatically label a large number of training data, we need to solve the following three challenges. (1) How to collect a large number of event pairs with a high probability of causality and employ them to label training data. (2) How to handle noisy distantly labeled sentences that do not have well-formed textual expressions to express causal semantics. (3) How to make better use of distantly labeled sentences for training. To this end, we firstly design a **Lexi**con Enhanced **Anno**tator (**LexiAnno**) to extract a large number of event pairs with a high probability of causality based on lexical knowledge and employ them to automatically label sentences via distant supervision. Secondly, we propose a **Common**sense **Filter** (**CommonFilter**) to refine distantly

Proceedings of the 28th International Conference on Computational Linguistics, pages 1544–1550
Barcelona, Spain (Online), December 8-13, 2020

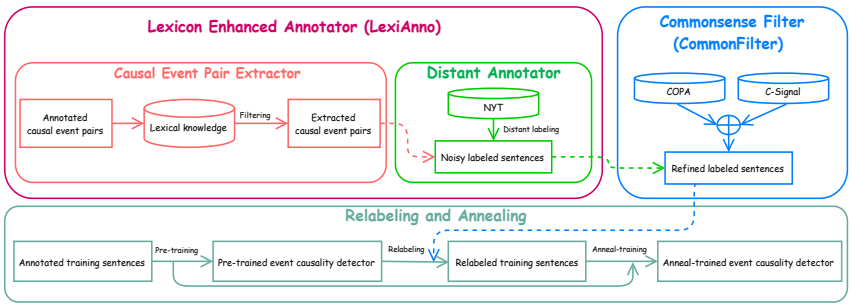

Figure 1: Overview of Knowledge Enhanced Distant Data Augmentation framework (KnowDis).

labeled sentences assisted with causal commonsense knowledge which makes them more well-formed to express the causal semantics. Thirdly, we employ **Relabeling** and **Annealing** strategies to make better use of distantly labeled sentences for training. Finally, we evaluate KnowDis on two datasets and achieve the best performance training with distantly labeled sentences on ECD. The following sections describe the architecture (Section 2) of KnowDis and the experimental results (Section 3) on the ECD task.

2 KnowDis

As shown in Figure 1, we illustrate the three main components of our proposed KnowDis in this section.

2.1 Lexicon Enhanced Annotator (LexiAnno)

LexiAnno aims to extract a large number of event pairs with a high probability from external lexicons based on the annotated causal event pairs via a *Causal Event Pair Extractor*, and employ them to collect preliminary noisy labeled sentences from external documents via a *Distant Annotator*.

Knowledge	How to extract	Why causality	Abbr.
WordNet	1) Extracting the synonyms and hypernyms from WordNet of head word of each event in e_{ij}. 2) Assembling the items from the two groups of two events to generate causal event pair set.	Items in each group are the synonyms and hypernyms of the original causal event pairs.	E^{wn}
VerbNet	1) Extracting the words from VerbNet under the same class as head word of each event in e_{ij}. 2) Assembling the items from the two groups of two events to generate causal event pair set.	Items in each group are in the same class of the original causal event pairs.	E^{vn}

Table 1: Extracting causal event pairs from lexical knowledge bases.

Causal Event Pair Extractor. We expand each event pair e_{ij} in annotated causal event pair set E^g via external dictionaries.[1]. Table 1 illustrates the details of how to extract E^{wn} and E^{vn} from WordNet (Miller, 1995) and VerbNet (Schuler, 2005). Eventually, we construct a filter via transE (Bordes et al., 2013) based on maximum interval method: $L = \sum_{(e_i,e_j)\in S} \sum_{(e'_i,e'_j)\in S'} [\lambda + d(\boldsymbol{e'_i}, \boldsymbol{e'_j}) - d(\boldsymbol{e_i}, \boldsymbol{e_j})]_+$ to sort extracted event pairs in ascending order of their distance and pick out the ones at the top of them with a high probability of causality, where S and S' are the causal and non-causal event pair set respectively.

Distant Annotator. We keep the top 10% sorted extracted event pairs to obtain E^f with a high probability of causality. Then we automatically label the 5% randomly selected sentences from NYT corpus[2] which contain any event pair e_{ij} in E^g and E^f as the noisy distantly labeled training data $\boldsymbol{D_n}$.

2.2 Commonsense Filter (CommonFilter)

CommonFilter aims to refine $\boldsymbol{D_n}$ assisted with causal commonsense knowledge to pick out labeled sentences which express causal semantics between events. Inspired by Luo et al. (2016), we introduce Pointwise Mutual Information (PMI) statistics (Church and Hanks, 1989) to indicate the causal semantics assisted with a *if-then* reasoning data Choice of Plausible Alternatives (COPA) (Gordon et al., 2011) and causal connectives. As shown in Table 2, we employ causality co-occurrences (f) of each word pair between cause-related (T_c) and effect-related (T_e) text and incorporate *necessity causality* (CS_{nec}) with

[1] *WordNet*: https://wordnet.princeton.edu/ and *VerbNet*: http://verbs.colorado.edu/ mpalmer/projects/verbnet.html

[2] NYT corpus is very large, so we randomly select 5% sentences from NYT corpus as the sentences to be labeled.

Source	Data Form	Cause-related Text	Effect-related Text
COPA	*Premise*: The woman hired a lawyer. *Alternative1*: She decided to sue her employer. (✓) *Alternative2*: She decided to run for office. (✗)	She decided to sue her employer	The woman hired a lawyer
Annotated Data	Kimani Gary, a young man who likes football, was killed in a police attack shortly after a tight match.	a police attack shortly after a tight match	Kimani Gary, a young man who likes football, was killed

Table 2: Cause-related and effect-related text from COPA and annotated data.

sufficiency causality (CS_{suf}) to model causal relation. Specifically, we calculate the CS_{nec} and CS_{suf} of each word pair (i_c, j_e) in (T_c, T_e) from COPA and annotated data, and causality score CS_s of two text spans (SP_1, SP_2) of each sentence s in $\boldsymbol{D_n}$ divided with connectives between two events:

$$CS_{nec}(i_c, j_e) = \frac{p(i_c|j_e)}{p^{\alpha}(i_c)} = \frac{p(i_c, j_e)}{p^{\alpha}(i_c)p(j_e)}, CS_{suf}(i_c, j_e) = \frac{p(j_e|i_c)}{p^{\alpha}(j_e)} = \frac{p(i_c, j_e)}{p(i_c)p^{\alpha}(j_e)} \tag{1}$$

$$p(i_c) = \frac{\sum_{w \in W} f(i_c, w_e)}{M}, p(j_e) = \frac{\sum_{w \in W} f(w_c, j_e)}{M}, p(i_c, j_e) = \frac{f(i_c, j_e)}{N}, M = \sum_{u \in W} \sum_{v \in W} f(u_c, v_e), \tag{2}$$

$$CS(i_c, j_e) = CS_{nec}(i_c, j_e)^{\lambda} CS_{suf}(i_c, j_e)^{1-\lambda}, CS_s(SP_1, SP_2) = \frac{1}{|SP_1| + |SP_2|} \sum_{i \in SP_1} \sum_{j \in SP_2} CS(i, j) \tag{3}$$

where, N is the size of all (T_c, T_e) pairs, W is all calculated words and α is a penalty value to penalize high-frequency words. Next, we sort and divide sentences in $\boldsymbol{D_n}$ into two parts based on CS_s, $\boldsymbol{D_n^c}$ in which the two events are connected by a causal connective from $\boldsymbol{C_{signal}}$ extracted from FrameNet (Baker et al., 1998) and PDTB2 (Group and others, 2008), and the $\boldsymbol{D_n^{nc}}$ in which are not. Finally, we keep the top 50% data in $\boldsymbol{D_n^c}$ and 10% data in $\boldsymbol{D_n^{nc}}$ as refined distantly labeled training data $\boldsymbol{D_r}$.

2.3 Relabeling and Annealing

Event Causality Detector. We formulate event causality detection as a sentence-level binary classification problem. Specifically, we design a binary classifier based on BERT (Devlin et al., 2019) to construct the *Event Causality Detector*. The input of the detector is the event pair e_{ij} and its corresponding sentence. We convert the sentence into BERT's input form, i.e. the sum of WordPiece embedding (Wu et al., 2016), position embedding, and segment embedding. We get the event representation $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ encoded by BERT. Then, we take the stitching of manual designed feature vector (same lexical, causal potential, and syntactic features representation as Gao et al. (Gao et al., 2019)) $\boldsymbol{f}$, $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ as the input of top MLP classifier. Finally, the output is a binary vector to indicate the causality of the input event pair e_{ij}.

We employ relabeling and annealing strategies to make better use of distantly labeled data for training. (1) *Relabeling*: We pre-train a detector on annotated data and employ it to relabel the refined distantly labeled training data $\boldsymbol{D_r}$ via self-training (Asai and Hajishirzi, 2020). Then, we collect the sentences that are relabeled as causal sentences to obtain the distantly relabeled training data $\boldsymbol{D_{rr}}$ which are more casual and informative for the training of ECD task. (2) *Annealing*: Distantly labeled training data may not be appropriate at the beginning of training for building an effective detector due to noises. Therefore, we employ the annealing training strategy (Kirkpatrick et al., 1983) to maximize the effectiveness of distantly labeled training data. In the beginning, we only employ annotated data for training, and with the increase of epochs, we added $\boldsymbol{D_{rr}}$ for training incrementally in a proportion of β.

3 Experiments

Datasets. (1) **ESC**: We use the same way to partition dataset as the SOTA method on ESC (Gao et al., 2019). Same as it, we use the last two topics as a development set. (2) **Causal-TB**: This dataset only contains 318 causal links which can further prove effectiveness of the proposed framework for solving the problem of data lacking. We use the same development set as ESC because of the SOTA method on this dataset (Mirza and Tonelli, 2014) does not partition the development set. Specifically, we conduct 5-fold cross-validation on the two datasets[3]. We tune the augmented proportion, α, and β on the development set. All the results are the average of three independent experiments.

[3]For each fold, we add extra distantly labeled data based on the annotated event pairs corresponding to this fold for training.

Parameters Setting. We apply the *base-uncase-bert* as the pre-trained BERT model. We set the learning rate of detector as 1e-5. Specifically, the dimension of the causal semantic space is 100. We set the α and β as 0.5 and 0.1 respectively based on the development set. We apply the early stop strategy and the SGD gradient strategy to optimize all models. We adopt *Precision* (P), *Recall* (R), *F1 vaule* (F1) as the evaluation metrics.

Compared Methods. We evaluate the performance of ECD on the same EventStoryLine corpus v0.9 (ESC) (Caselli and Vossen, 2017) and Causal-TimeBank (Causal-TB) (Mirza and Tonelli, 2014) dataset as SOTA methods. We select some typical methods and SOTA methods on ESC and Causal-TB respectively to make comparisons: (1) **Cheng et al. (2017)** and **Choubey et al. (2017)**: two dependency path based sequential neural models which have shown effectiveness on ESC. (2) **Gao et al. (2019)**: the SOTA method which models the document-level structures for event causality detection on ESC. (3) **Mirza et al. (2014)**: a strong supervised classifier with gold causal signals on Causal-TB. (4) **BERT**: our proposed detector (2.3), a strong baseline for comparison which performs significantly well in classification tasks. The other compared methods are not open source entirely, so we construct a BERT-based model as a strong baseline to verify our data augmentation framework. (5) **EDA**: training detector (2.3) with extra data augmented by EDA which is a easy data augmentation framework (Wei and Zou, 2019). We also employ relabeling and annealing strategy when training with EDA. Finally, we automatically label 10132 sentences via KnowDis. We sample 100 sentences for manual evaluation, 82% of which clearly express the causal semantics (3 assessors, Cohen's kappa = 0.88).

3.1 Comparisons with SOTA Methods on Event Causality Detection

ESC				
Methods	**P**	**R**	**F1**	**∇**
Cheng et al.(2017)	34.0	41.5	37.4	-
Choubey et al.(2017)	32.7	44.9	37.8	-
Gao et al.(2019)	37.4	55.8	44.7	-
BERT	36.6	59.7	45.3	+0.6
EDA(Wei and Zou, 2019)	38.8	62.9	48.0	+3.3
KnowDis (our method)	39.7	66.5	**49.7**	+5.0

Table 3: Performance of compared methods for ECD on ESC. ∇ means the points higher than the SOTA method on F1 value.

Causal-TB				
Methods	**P**	**R**	**F1**	**∇**
Mirza et al. (2014)	74.6	35.2	47.8	-
BERT	39.0	60.5	47.4	-0.4
EDA(Wei and Zou, 2019)	40.2	61.2	48.5	+0.7
KnowDis (our method)	42.3	60.5	**49.8**	+2.0

Table 4: Performance of compared methods for ECD on Causal-TB. ∇ means the points higher than the SOTA method on F1 value.

Table 3 and 4 shows the results of our model compared with SOTA methods. From the results, we could have the main following observations. (1) ***Effectiveness of our method***: Our method (**KnowDis**) significantly improves the performance of ECD by 5.0 and 2.0 points on F1 value on two datasets respectively. It illustrates that the augmented training data labeled via distant supervision, and refined via causal commonsense knowledge can provide more effective assistance for ECD task. (2) ***Necessity of causal-related knowledge***: We can observe that training with extra data augmented with **EDA** and **KnowDis** can both improve the performance of ECD task which shows that more training data can introduce more causal knowledge to alleviate data scarcity. However, the sentences produced via **EDA** are not refined by causal-related knowledge such as causal lexical and causal commonsense knowledge. Compared to it, the further significant improvement of **KnowDis** proves the necessity of the causal-related knowledge, and also illustrates that our model can produce more suitable training data for the ECD task.

3.2 Effectiveness of Main Components on Event Causality Detection

Method	**P**	**R**	**F**	**∇**
BERT (Baseline)	36.6	59.7	45.3	-
+Design features	36.8	63.0	46.5	+1.2
+Annotated causal ep.†	37.5	67.0	48.1	+2.8
+Extracted causal ep.†	39.7	66.5	**49.7**	+4.4

Table 5: Effectiveness of LexiAnno (2.1) and Detector (2.3) on ESC.

Method	**P**	**R**	**F**	**∇**
KnowDis (Our model)	39.7	66.5	**49.7**	-
-Causal connective	37.3	67.9	48.1	-1.6
-Causality co-occurrence	36.7	67.2	47.5	-2.2
-Relabeling	38.1	67.0	48.6	-1.1
-Annealing	37.6	67.8	48.4	-1.3

Table 6: Effectiveness of CommonFilter (2.2), Relabeling and Annealing (2.3) on ESC.

Table 5 and 6 tries to show the effectiveness of the key parts of our method on event causality detection (ECD). † denotes the same filtering and training processes except that the causal event pairs employed for distant labeling are different. From the results, we could have the following observations. (1) ***Effectiveness of distant labeling***: The results of **Annotated causal ep.†** and **+Extracted causal ep.†** illustrates that the distantly labeled augmented training data can effectively alleviate the problem of data scarcity on ECD task. (2.1) (2) ***Effectiveness of LexiAnno***: Compared to sentences labeled only based on annotated causal event pairs (**+Annotated causal ep.†**), training with sentences labeled based on extracted causal event pairs from knowledge bases (**+Extracted causal ep.†**) can bring effective and diverse knowledge for understanding event-causal semantics. (2.1) (3) ***Effectiveness of CommonFilter***: The results of **-Causal connective** and **-Causality co-occurrence** show that our proposed commonsense filter (2.2) which introduces causal commonsense knowledge to refine distantly labeled training data can effectively enhance the causal-related semantics of augmented data. Specifically, the causal commonsense knowledge is more useful than causal connective knowledge because the former is more extensive than the latter in the expression of cause and effect. (4) ***Effectiveness of Relabeling***: The results of **-Relabeling** show that relabeling (2.3) can reduce the noisy of distantly labeled data. (5) ***Effectiveness of Annealing***: The results of **-Annealing** show that the annealing (2.3) can make better use of noisy distantly labeled data. Relabeling and annealing can both be applied to other distant supervision tasks.

3.2.1 Effectiveness of Different Proportion of Distantly Labeled Data

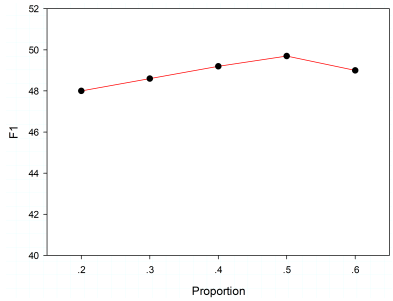

Figure 2: Effectiveness of different proportion of retained distantly labeled data in $\boldsymbol{D_n^c}$ on ESC.

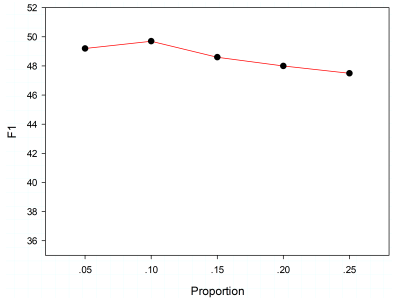

Figure 3: Effectiveness of different proportion of retained distantly labeled data in $\boldsymbol{D_n^{nc}}$ on ESC.

Figure 2 and 3 tries to show the effectiveness of different proportion of retained distantly labeled data in $\boldsymbol{D_n^c}$ and $\boldsymbol{D_n^{nc}}$. From the results, we could have the main following observations. (1) The more data retained of the distant label data in $\boldsymbol{D_n^c}$, the more effective knowledge can be brought for training. However, when the retained data exceeds 50%, the noise caused by $\boldsymbol{D_n^c}$ is greater than the impact of effective knowledge. (2) Introducing appropriate distant label data in $\boldsymbol{D_n^{nc}}$ can bring additional effective knowledge but it contains more harmful noise than the data in $\boldsymbol{D_n^c}$.

4 Conclusion

In this paper, we try to employ distant supervision to alleviate the data lacking problem on causal-related task. We propose a knowledge enhanced distant data augmentation framework (KnowDis) for event causality detection. Our method achieves the SOTA performance on EventStoryLine corpus and Causal-TimeBank dataset assisted with knowledge enhanced distantly labeled training data. In the future, we will introduce more causal-related resources and apply KnowDis for other relational tasks.

Acknowledgements

This work is supported by the National Key R&D Program of China (2020AAA0106400), the National Natural Science Foundation of China (No.61533018, No. 61922085, No.61806201) and the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006). This work is also supported by a grant from Ant Group and Beijing Academy of Artificial Intelligence (BAAI).

References

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. *ArXiv*, abs/2004.10157.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *COLING-ACL*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

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.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–419, Vancouver, Canada, July. 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.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. *arXiv preprint arXiv:1707.07343*.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *ACL*.

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, June. 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.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2011. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *SemEval@NAACL-HLT*.

PDTB Research Group et al. 2008. The pdtb 2.0. *Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania*.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997.

Zhichao Hu and Marilyn A Walker. 2017. Inferring narrative causality between event pairs in films. *arXiv preprint arXiv:1708.09496*.

Scott Kirkpatrick, C. D. Gelatt, and Mario P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220 4598:671–80.

Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *KR*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL/IJCNLP*.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *COLING*.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *The 26th international conference on computational linguistics*, pages 64–75. ACL.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra-and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1733–1743.

Jong-Hoon Oh, Kentaro Torisawa, Canasai Kruengkrai, Ryu Iida, and Julien Kloetzer. 2017. Multi-column convolutional neural networks with causality-attention for why-question answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 415–424. ACM.

Mehwish Riaz and Roxana Girju. 2014. Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 48–57.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. Towards causal explanation detection with pyramid salient-aware network.