Transfer Learning for Causal Sentence Detection

Manolis Kyriakakis[1], Ion Androutsopoulos[2], Joan Ginés i Ametllé[1], Artur Saudabayev [1]
[1]Causaly, London, UK
[2]Department of Informatics, Athens University of Economics and Business, Greece
`{m.kyriakakis,joan.g,artur}@causaly.com, ion@aueb.gr`

Abstract

We consider the task of detecting sentences that express causality, as a step towards mining causal relations from texts. To bypass the scarcity of causal instances in relation extraction datasets, we exploit transfer learning, namely ELMO and BERT, using a bidirectional GRU with self-attention (BIGRUATT) as a baseline. We experiment with both generic public relation extraction datasets and a new biomedical causal sentence detection dataset, a subset of which we make publicly available. We find that transfer learning helps only in very small datasets. With larger datasets, BIGRU-ATT reaches a performance plateau, then larger datasets and transfer learning do not help.

1 Introduction

A wide range of biomedical questions, from what causes a disease to what drug dosages should be recommended and which side effects might be triggered, center around detecting particular causal relationships between biomedical entities. Causality, therefore, has long been a focus of biomedical research, e.g., in medical diagnostics (Rizzi, 1994), pharmacovigilance (Agbabiaka et al., 2008), and epidemiology (Karhausen, 2000). The most common way to detect causal relationships is by carrying highly controlled randomized controlled trials, but it is also possible to mine evidence from observational studies and meta-analyses (Ward and Johnson, 2008), where information is often expressed in natural language (e.g., journal articles or clinical study reports).

In natural language processing (NLP), causality detection is often viewed as a type of relation extraction, where the goal is to determine which relations (e.g., part-whole, content-container, cause-effect), if any, hold between two entities in a text (Hendrickx et al., 2009), using deep learning in most recent works (Bekoulis et al., 2018; Zhang et al., 2018). The same view of causality detection is typically adopted in biomedical NLP (Cohen and Demner-Fushman, 2014; Li and Mao, 2019).

Existing relation extraction datasets, however, contain few causal instances, which may not allow relation extraction methods to learn to infer causality reliably. Note that causality can be expressed in many ways, from using explicit lexical markers (e.g., "smoking *causes* cancer") to markers that do not always express causality (e.g., "heavy smoking *led* to cancer" vs. "the nurse *led* the patient to her room") to no explicit markers ("she was infected by a virus and admitted to a hospital"). Also, existing relation extraction datasets contain sentences from generic, not biomedical documents.

In this paper, we focus on *detecting causal sentences*, i.e., sentences conveying at least one causal relation. This is a first step towards mining causal relations from texts. Once causal sentences have been detected, computationally more intensive relation extraction methods can be used to identify the exact entities that participate in the causal relations and their roles (cause, effect). To bypass the scarcity of causal instances in relation extraction datasets, we exploit *transfer learning*, namely ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018), comparing against a bidirectional GRU with self-attention (Cho et al., 2014; Bahdanau et al., 2015). We experiment with *generic* public relation extraction datasets and a new larger *biomedical* causal sentence detection dataset, a subset of which we make publicly available.[1]

Unlike recently reported results in other NLP tasks (Peters et al., 2018; Devlin et al., 2018; Peters et al., 2019), we find that transfer learning

---

[1]We cannot provide the entire biomedical dataset, because it is used to develop commercial products. We report, however, results for both the entire biomedical dataset and the publicly available subset.

helps only in datasets with hundreds of training instances. When a few thousands of training instances are available, BIGRUATT reaches a performance plateau (both in generic and biomedical texts), then increasing the size of the dataset or employing transfer learning does not improve performance. We believe this is the first work to (a) focus on causal sentence detection as a binary classification task, (b) consider causal sentence detection in both generic and biomedical texts, and (c) explore the effect of transfer learning in this task.

2 Methods

BIGRUATT: Our baseline model is a bidirectional GRU (BIGRU) with self-attention (BIGRUATT) (Cho et al., 2014; Bahdanau et al., 2015), a classifier that has been reported to perform well in short text classification (Pavlopoulos et al., 2017; Chalkidis et al., 2019). The model views each sentence as the sequence $\langle e_1, \ldots, e_n \rangle$ of its word embeddings (Fig. 1). We use WORD2VEC embeddings (Mikolov et al., 2013) pre-trained on approx. (a) 3.5 billion tokens from PUBMED texts (McDonald et al., 2018)[2] or (b) 100 billion tokens from Google News.[3] The BIGRU computes two lists H_f, H_b of hidden states, reading the word embeddings left to right and right to left, respectively. The corresponding elements of H_f, H_b are then concatenated to form the output H of the BIGRU:

$$
\begin{aligned}
H^f &= \left\langle h_1^f, \ldots, h_n^f \right\rangle = \text{GRU}^f(e_1, \ldots, e_n) \\
H^b &= \left\langle h_1^b, \ldots, h_n^b \right\rangle = \text{GRU}^b(e_1, \ldots, e_n) \\
H &= \left\langle [h_1^f; h_1^b], \ldots, [h_n^f; h_n^b] \right\rangle
\end{aligned}
$$

where ^{f}, ^{b} indicate the forward and backward directions, $e_i \in \mathbb{R}^{d_e}$, $h_i^f, h_i^b \in \mathbb{R}^{d_h}$, and ';' denotes concatenation.[4] A linear attention computes an attention score $a_i \in \mathbb{R}$ for each element h_i of H:

$$
\widetilde{a}_i = u_{att} \cdot h_i, \quad a_i = \text{softmax}(\widetilde{a}_i; \widetilde{a}_1, \ldots, \widetilde{a}_n)
$$

where $u_{att} \in \mathbb{R}^{2 \times d_h}$ and $\cdot$ is the dot product. A sentence embedding s, representing the entire sentence, is then formed as the weighted (by the attention scores) sum of the elements of H and is

Figure 1: Illustration of the BIGRUATT model.

passed to a logistic regression (LR) layer to estimate the probability p that a sentence is causal:

$$
s = \sum_{i=1}^{n} a_i h_i, \quad p = \sigma(u_p \cdot s + b_p)
$$

where $u_p \in \mathbb{R}^{2 \times d_h}$, $b_p \in \mathbb{R}$, and σ is the sigmoid function. We use cross-entropy loss, the Adam optimizer (Kingma and Ba, 2015), and dropout layers (Srivastava et al., 2014) before and after the BIGRU (Fig. 1). Word embeddings are not updated.

BIGRUATT+ELMO: ELMO (Peters et al., 2018) produces word embeddings by passing the input text (in our case, a sentence) to a pre-trained stacked bidirectional LSTM language model (LM). It then uses a linear combination of the states of the LM (from the different layers of the stacked LSTM) at each word position to produce the corresponding word embedding. Like WORD2VEC, ELMO (its LM) is pre-trained on large corpora. However, ELMO maps occurrences of the same word to possibly different embeddings, depending on context. Furthermore, it uses CNNs (LeCun et al., 1989) to produce the initial word embeddings (that are fed to the LM) from word characters, alleviating the problem of out-of-vocabulary words. BIGRUATT+ELMO is the same as BIGRUATT, except that the embedding of each word is now the concatenation of its WORD2VEC and ELMO embeddings. We do not update the parameters of ELMO and the word embeddings when training BIGRUATT+ELMO. We used the original pre-trained ELMO of Peters et al. (2018).[5] For biomedical sentences, we also experimented with an ELMO model pre-trained on PUBMED texts, but performance was very similar as with the original ELMO.

BERT+LR: BERT (Devlin et al., 2018) is a model based on Transformers (Vaswani et al., 2017), pre-

[4]In our experiments, $d_h = 128$; d_e is 300 for Google News and 200 for biomedical embeddings; d_e increases by 1,024 when ELMO is added. We also tried LSTMs (Hochreiter and Schmidhuber, 1997), but performance was similar.
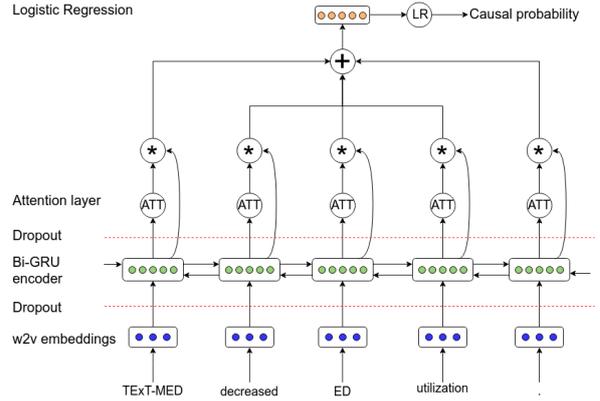
trained on large corpora to predict (a) masked words from their left and right contexts, and (b) the next sentence. For a new NLP task, a task-specific layer is added on top of a pre-trained BERT model. The extra layer is trained jointly with BERT on task-specific data (in our case, a causal sentence detection dataset), a process that also fine-tunes the parameters of the pre-trained BERT for the new task. In BERT+LR, we add a logistic regression (LR) layer on top of BERT, which estimates the probability that the input sentence is causal. The LR layer is fed with the embedding of the 'classification' token, which BERT also produces for each sentence. We used the pre-trained 'base' BERT model of Devlin et al. (2018), which we fine-tuned jointly with the LR layer. For biomedical sentences, we also experimented with BIOBERT (Lee et al., 2019), a BERT model pre-trained on biomedical texts, but performance was very similar.

BERT+BIGRUATT: Common practice (Devlin et al., 2018; Peters et al., 2019) is to combine BERT with very shallow task-specific models, usually only an LR layer. To explore if deeper task-specific models can yield improved performance, we replaced the LR layer of BERT+LR with BIGRUATT, leading to BERT+BIGRUATT. This is the same as BIGRUATT, but uses the context-aware word embeddings that BERT produces at its top layer as the input to BIGRUATT, instead of WORD2VEC embeddings. Again, we use the 'base' pre-trained BERT model of (Devlin et al., 2018), and we fine-tune the entire BERT+BIGRUATT network on causal sentence detection datasets.

LR (n-grams): A plain LR classifier with TF-IDF n-gram features (word n-grams, $n = 1, 2, 3$).[6]

3 Datasets

SemEval-2010 (Task 8): This dataset contains 10,674 samples, of which 1,325 causal (Hendrickx et al., 2009). Each sample is a sentence annotated with a pair of entities and the type of their relationship. Since we are only interested in causality, we treat sentences with a *Cause-Effect* relationship as causal, and all the others as non-causal.

Causal-TimeBank (CausalTB): In this dataset (Mirza et al., 2014), we identified causal sentences using C-SIGNAL (causal signal) and CLINK (causal

link) tags, discarding causal relationships between entities from different sentences, following Li and Mao (2019). The resulting dataset contains 2,470 sentences, of which 244 are causal.

Event StoryLine (EventSL): In this dataset (Caselli and Vossen, 2017), we detected causal sentences by examining the CAUSES and CAUSED_BY attributes in the PLOT_LINK tags, again following Li and Mao (2019). Again, we discarded causal relationships between entities from different sentences. The resulting dataset contains 4,107 sentences, of which 77 are causal.

BioCausal: The full biomedical causal detection dataset we developed (**BioCausal-Large**) contains 13,342 sentences from PUBMED, of which 7,562 causal. Each sentence was annotated by a single annotator familiar with biomedical texts.[7] The publicly available subset (**BioCausal-Small**) contains 2,000 sentences, of which 1,113 causal.[8]

Li and Mao (2019) report that SemEval-2010 contains a large number of causal samples with explicit causal markers; by contrast, CausalTB and EventSL contain more complex causal relations with no explicit clues. BioCausal includes causal sentences both with and without explicit clues.

SemEval-2010, CausalTB, and EventSL are highly imbalanced, with the vast majority of sentences being non-causal. To prevent a high bias towards the non-causal class, in our experiments we randomly selected 2500, 500, 200 non-causal sentences respectively, discarding the rest. The resulting causal to non-causal ratios (Table 1) are, thus, roughly 1:2 (SemEval, CausalTB) or 1:3 (EventSL). By contrast, the BioCausal (Large and Small) datasets are roughly balanced. All five datasets were then split into train (70%), validation (15%) and test (15%) subsets, maintaining the same ratio between the two classes in the three subsets.

4 Experimental results

Tables 1–2 report our experimental results. For each neural model, we performed 10 repetitions (with different random seeds) and report averages and standard deviations. For completeness, we show precision, recall, F1, and area under the precision-recall curve (AUC), though AUC scores are the main ones to consider, since they examine

---

[6]We used the LR code of SCIKIT-LEARN (https://scikit-learn.org/). For all other methods, we used our own PyTorch implementations (https://pytorch.org/), with the BERT API of https://github.com/huggingface/pytorch-pretrained-BERT.

[7]The average inter-annotator agreement on a sample of 300 sentences was 79.36%. Cohen's Kappa was 0.56.

[8]BioCausal-Small is available at https://archive.org/details/CausalySmall.

Dataset (causal:non-causal)	SemEval (1,325 : 2,500)				CausalTB (244 : 500)				EventSL (77 : 200)				BioCausal-Small (1,113 : 887)			
Model	F1	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC
LR (n-grams)	76.22	88.67	66.83	87.50	36.36	**100.00**	22.22	65.02	42.86	**100.00**	27.27	73.55	77.49	73.91	81.44	87.65
BIGRUATT	90.64 ±0.70	93.96 ±1.71	87.59 ±1.52	96.57 ±0.32	69.98 ±3.58	67.04 ±5.16	73.89 ±6.60	74.38 ±4.16	63.65 ±10.12	70.09 ±7.47	60.91 ±17.28	70.36 ±9.84	85.97 ±0.90	83.57 ±2.03	88.62 ±2.69	93.91 ±0.88
BIGRUATT+ELMO	**92.81** ±0.78	**94.45** ±0.94	91.26 ±1.77	97.03 ±1.44	75.08 ±4.20	81.29 ±5.43	70.28 ±6.81	82.06* ±3.59	66.55 ±7.82	77.47 ±5.05	59.09 ±10.17	77.31 ±4.84	**87.32** ±0.78	**89.46** ±2.34	85.39 ±2.50	**94.95** ±0.33
BERT+LR	91.55 ±0.53	86.62 ±1.16	**97.09** ±0.67	96.94 ±2.25	**80.55** ±3.62	71.17 ±6.02	**93.33** ±3.33	82.26* ±3.41	72.35 ±5.36	62.44 ±8.21	**87.17** ±1.16	78.15* ±4.58	85.64 ±0.61	78.87 ±1.16	**93.71** ±1.54	90.75 ±3.69
BERT+BIGRUATT	91.45 ±0.59	86.80 ±1.28	96.63 ±0.60	**97.61** ±0.29	80.06 ±2.94	74.52 ±4.46	86.94 ±5.56	**84.27*** ±1.71	**73.09** ±5.27	66.15 ±7.86	83.64 ±10.60	**84.17*** ±4.04	85.87 ±0.88	79.43 ±1.53	93.47 ±1.08	93.75 ±0.45

Table 1: Precision, recall, F1, AUC on the four publicly available datasets, averaged over 10 repetitions, with standard deviations (±). Next to each dataset name, we show in brackets the total causal and non-causal sentences that we used. The best results are shown in bold. The best AUC results are also shown in gray background. In the AUC columns, stars indicate statistically significant ($p \leq 0.05$) differences compared to BIGRUATT.

performance at multiple classification thresholds; the other measures are computed only for a particular threshold, which was 0.5 in our experiments.

Model	F1	P	R	AUC
LR (n-grams)	79.21	76.75	81.82	86.54
BIGRUATT	85.84 ±0.36	84.28 ±0.66	87.47 ±0.75	93.71 ±0.15
BIGRUATT+ELMO	86.77 ±0.52	**87.46** ±1.29	86.12 ±1.49	94.64* ±0.26
BERT+LR	**87.33** ±0.47	82.11 ±0.89	**93.27** ±0.67	92.77 ±1.57
BERT+BIGRUATT	87.09 ±0.34	81.70 ±0.58	93.25 ±0.40	**94.70** ±0.24

Table 2: Results on BioCausal-Large (7,562 : 5,780).

Focusing on AUC scores, BIGRUATT outperforms the simpler LR with n-grams by a wide margin, with the exception of EventSL, which is probably too small for the capacity of BIGRU-ATT.[9] The precision of LR is perfect on CausalTB and EventSL, at the expense of very low recall, suggesting that LR learned perfectly few high-precision n-grams in those datasets. Transfer learning (ELMO, BERT) improves the AUC of BIGRUATT by a wide margin in the two smallest datasets (CausalTB, EventSL), which contain only hundreds of instances, and the AUC differences from BIGRUATT are statistically significant (stars in Table 1), except for BIGRUATT+ELMO in EventSL.[10] However, in the other three datasets which contain thousands of instances, the AUC differences between transfer learning and plain BIGRUATT are small, with no statistical significance in most cases. Also, the AUC scores of all methods on BioCausal-Large are close to those on BioCausal-Small, despite the fact that BioCausal-Large is approx. seven times larger. Similar observations can be made by looking at the F1 scores.

[9] Indeed BIGRUATT overfits the training set of EventSL.

[10] We performed two-tailed Approximate Randomization tests (Dror et al., 2018), $p \leq 0.05$, with 10k iterations, randomly swapping in each iteration 50% of the decisions (over all tested sentences) across the two compared methods. When testing statistical significance, for each method we use the repetition (among the 10) with the best validation F1 score.

It seems that causal sentence detection, at least with the neural methods we considered, reaches a plateau with few thousands of training sentences both in generic and biomedical texts; then increasing the dataset size or employing transfer learning does not help. The latter finding is not in line with previously reported results (Peters et al., 2018; Devlin et al., 2018; Peters et al., 2019), where ELMO and BERT were found to improve performance in most NLP tasks without studying, however, the effect of dataset size. Furthermore, BERT+BIGRUATT consistently performed better than BERT+LR in AUC (but not in F1), which casts doubts on the practice of adding only shallow task-specific models to BERT.[11] BIGRUATT+ELMO is competitive in AUC to BERT+BIGRUATT (with the exception of EventSL). Mainly comprised of sentences with simple explicit causal statements (Li and Mao, 2019), SemEval expectedly demonstrated the best classification performance across datasets.

5 Related and Future Work

Recent work on (causal) relation extraction uses LSTMs (Zhang et al., 2017) or CNNs (Li and Mao, 2019), assuming however that the spans of the two entities (cause, effect) are known. A notable exception is the model of Bekoulis et al. (2018), which jointly infers the spans of the entities and their relationships. Such finer relation extraction methods, however, are computationally more expensive than our causal sentence detection methods, especially when they involve parsing (Zhang et al., 2018). We plan to consider pipelines where computationally cheaper causal sentence detection components will first detect sentences likely to express causality, and then finer relation extraction components will pinpoint the entities, the type of

[11] We note, however, that the AUC difference between BERT+LR and BERT+BIGRUATT is statistically significant ($p \leq 0.05$) only in BioCausal-Large.

causality (e.g., up-regulate), and entity roles.

Appendix

A Hyper-parameters

Batch sizes of 128, 32, 16, 32 and 256 were used for Semeval, CausalTB, EventSL, BioCausal-Small and BioCausal-Large, respectively, for all neural models. Trainable parameters were initialized using the default PyTorch initialization methods except from self-attention weights where the method of Glorot and Bengio (2010) was used.

BIGRU and BIGRUATT+ELMO were trained for 100 epochs using Adam (Kingma and Ba, 2015) with initial learning rate 2e^{-3} , $\beta_1/\beta_2 = 0.9/0.999$ and $eps = 1e^{-8}$. The learning rate was decayed linearly every 20 epochs as $lr_{new} \leftarrow lr_{prev} \cdot 0.75$. Gradients were clipped using a clip norm threshold of 0.25. The GRU's hidden size was set to 128 and its depth to 1. A dropout of 0.5 was applied to the input and output connections of the BIGRU encoder. Validation F1 was checked periodically in order to keep the model's checkpoint with the best validation performance.

BERT and BERT+BIGRUATT used the BERT-BASE uncased pre-trained model, which has 12 layers, 768 hidden size, 12 attention heads, and 110M parameters. For both models the entire network was fine-tuned for 10 epochs using Adam with a very small learning rate of 2e^{-5}, $\beta_1/\beta_2 = 0.9/0.999$, $eps = 1e^{-6}$, L2 weight decay of 0.01 and linear warmup of 0.1. A dropout of 0.1 was applied to all BERT-specific layers. For BERT+BIGRUATT, an additional dropout of 0.5 was applied to the input and output connections of its BIGRU encoder. Similarly to BIGRU and BIGRUATT+ELMO, the hidden size of the GRU was set to 128 and its depth to 1.

References

Taofikat B. Agbabiaka, Jelena Savović, and Edzard Ernst. 2008. Methods for causality assessment of adverse drug reactions. *Drug Safety*, 31(1):21–37.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*, San Diego, California.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium.

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.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification:a case study in EU legislation. In *NAACL Workshop in Natural Legal Language Processing*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar.

Kevin Bretonnel Cohen and Dina Demner-Fushman. 2014. *Biomedical Natural Language Processing*. John Benjamins.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1383–1392.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99, Boulder, Colorado.

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

Lucien R. Karhausen. 2000. Causation: The elusive grail of epidemiology. *Medicine, Health Care and Philosophy*, 3(1):59–67.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Yann LeCun, Bernhard E. Boser, John S. Denker, Don Henderson, R. E. Howard, W. Hubbard, and Larry Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Pengfei Li and Kezhi Mao. 2019. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115:512–523.

Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1849–1860, Brussels, Belgium.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark.

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

Matthew Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pre-trained representations to diverse tasks. *CoRR*, abs/1903.05987.

Dominick A. Rizzi. 1994. Causal reasoning and the diagnostic process. *Theoretical medicine*, 15:315–33.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Andrew Ward and Pamela Jo Johnson. 2008. Addressing confounding errors when using non-experimental, observational data to make causal claims. *Synthese*, 163(3):419–432.

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

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark.