Event Causality Recognition Exploiting Multiple Annotators' Judgments and Background Knowledge

Kazuma Kadowaki[§‡], **Ryu Iida**[§¶], **Kentaro Torisawa**[§¶],
Jong-Hoon Oh[§], and **Julien Kloetzer**[§]

[§]Data-driven Intelligent System Research Center (DIRECT),
National Institute of Information and Communications Technology (NICT)
[‡]Advanced Technology Laboratory, The Japan Research Institute, Limited
[¶]Graduate School of Science and Technology, Nara Institute of Science and Technology
{kadowaki,ryu.iida,torisawa,rovellia,julien}@nict.go.jp

Abstract

We propose new BERT-based methods for recognizing event causality such as "smoke cigarettes" → "die of lung cancer" written in web texts. In our methods, we grasp each annotator's policy by training multiple classifiers, each of which predicts the labels given by a single annotator, and combine the resulting classifiers' outputs to predict the final labels determined by majority vote. Furthermore, we investigate the effect of supplying background knowledge to our classifiers. Since BERT models are pre-trained with a large corpus, some sort of background knowledge for event causality may be learned during pre-training. Our experiments with a Japanese dataset suggest that this is actually the case: Performance improved when we pre-trained the BERT models with web texts containing a large number of event causalities instead of Wikipedia articles or randomly sampled web texts. However, this effect was limited. Therefore, we further improved performance by simply adding texts related to an input causality candidate as background knowledge to the input of the BERT models. We believe these findings indicate a promising future research direction.

1 Introduction

Event causality, such as "smoke cigarettes" → "die of lung cancer," is critical knowledge for NLP applications such as machine reading (Rajpurkar et al., 2016). For the task of recognizing event causality written in web texts, we propose new BERT-based methods that exploit independent labels in a gold dataset provided by multiple annotators. In the creation of the dataset we used (Hashimoto et al., 2014), three annotators independently labeled the data and the final labels were determined by majority vote. In the previous work, each annotator's independent judgments were ignored, but in our proposed method, we exploit each annotator's judgments in predicting the majority vote labels.

The dataset we used had a reasonable degree of inter-annotator agreement (Fleiss' Kappa value was 0.67), but a discrepancy remained among the annotators. Despite this discrepancy, we assume that their judgments are more or less consistent and that we can improve performance by training multiple classifiers, each from the labels provided by an individual annotator to grasp her/his policy, and by combining the resulting outputs of these classifiers.

Researchers have studied how to exploit the differences between the behaviors of annotators and crowd workers to improve the quality of gold datasets (Snow et al., 2008; Zaidan and Callison-Burch, 2011; Zhou et al., 2012; Jurgens, 2013; Plank et al., 2014; Jamison and Gurevych, 2015; Felt et al., 2016; Li et al., 2017). In an attempt that resembles ours, one study (Jamison and Gurevych, 2015) successfully improved the performance of several NLP tasks by computing the agreement ratio of each training instance and using only those instances with high agreement. Another work (Plank et al., 2014) improved part-of-speech tagging by measuring the inter-annotator agreement on a small number of sampled data and incorporating this value during training via a modified loss function. However, neither of them directly used each annotator's judgments, as we did in this work.

As another research direction, we also investigate how to appropriately exploit *background knowledge*. In previous work, text fragments such as binary patterns (e.g., "A causes B") and texts expressing causalities (e.g., "He died due to lung cancer") retrieved from large corpora were given to causality recognizers as background knowledge (Hashimoto et al., 2014; Kruengkrai et al., 2017), as well as association among words (Torisawa,

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5816–5822,
Hong Kong, China, November 3–7, 2019.

2006; Riaz and Girju, 2010; Do et al., 2011), semantic polarities (Hashimoto et al., 2012), answers obtained from a web-based open-domain why-QA system and other causality related texts (Kruengkrai et al., 2017), and causality-related word embeddings (Xie and Mu, 2019).

In this work, we investigate whether BERT (Devlin et al., 2019) (especially its pre-training) enables novel ways to exploit background knowledge. Our assumption is that if a BERT model is pre-trained using a large amount of causality-rich texts, it can learn some sort of background knowledge from the text. If the pre-training is adequately performed, background knowledge in the form of text fragments and a special mechanism for dealing with them might become obsolete. Our experimental results show that a BERT model pre-trained with causality-rich texts achieved significantly better performance than models using Wikipedia articles or randomly sampled web texts, both of which can be viewed as texts that do not specifically focus on causality. But the BERT model does not seem to sufficiently capture background knowledge, at least in our task setting. Further improvement is possible by simply concatenating, to an input causality candidate, text fragments related to it as background knowledge.

In our experiments, we show that our best method significantly outperformed a state-of-the-art method (Kruengkrai et al., 2017) by about 5% in average precision.

2 Proposed Method

We propose three BERT-based methods for event-causality recognition, and show an overview of them in Figure 1. All of the methods take input matrix $\mathbf{x}$, which represents an input causality candidate such as "smoke cigarettes" $\rightarrow$ "die of lung cancer," and obtain each annotator's labels (either *ProperCausality* or *NonProperCausality*), which are denoted by y^A, y^B, and y^C for three annotators A, B, and C, respectively. These are used for predicting final labels y^{MV}, which are determined by majority vote. Here, we assume that the dataset was labeled by three annotators, A, B and C, but extending the methods to deal with an arbitrary number of annotators is straightforward.

The proposed methods compute the probability $P(y^{\mathrm{MV}}|\mathbf{x})$ that a causality candidate represented by $\mathbf{x}$ expresses a proper event causality. We regard the candidate as proper if and only if $P(y^{\mathrm{MV}}|\mathbf{x}) > \theta$ ($\theta = 0.5$).

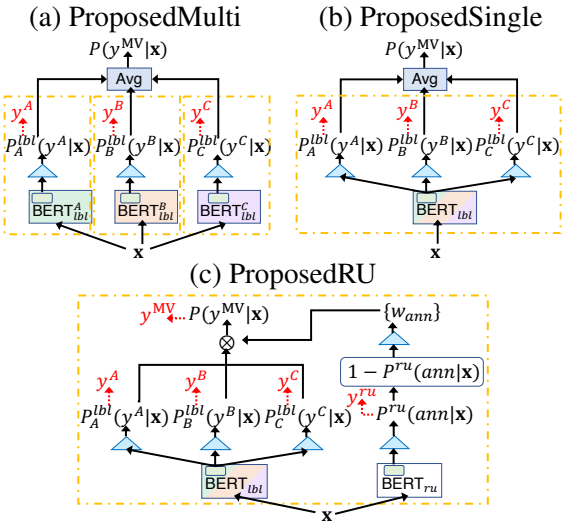


Figure 1: Proposed architectures, where red symbols (e.g., y^A) stand for gold labels used in training.

$$P(y^{\mathrm{MV}}|\mathbf{x}) = \sum_{ann} w_{ann} P^{lbl}_{ann}(y^{ann}|\mathbf{x}) \quad (1)$$

$$P^{lbl}_{ann}(y^{ann}|\mathbf{x}) = \mathrm{softmax}(\mathbf{W}_{ann}\mathrm{BERT}_{lbl}(\mathbf{x})) \quad (2)$$

$$P^{ru}(ann|\mathbf{x}) = \mathrm{softmax}(\mathbf{W}_D\mathrm{BERT}_{ru}(\mathbf{x})) \quad (3)$$

$$w_{ann} = \mathrm{softmax}_{ann}(1 - P^{ru}(ann|\mathbf{x})) \quad (4)$$

$\mathbf{W}_D \in \mathbb{R}^{3\times d}$ and $\mathbf{W}_{ann} \in \mathbb{R}^{2\times d}$ are trainable matrices, where d is size of hidden state in BERT models.

Table 1: Equations used in ProposedRU

ProposedMulti: The outputs of this method are achieved by an ensemble of three classifiers, each of which is independently trained (or, more precisely, fine-tuned) with the judgments of one of three annotators to mimic her/his judgments. The architecture of this method is shown in Figure 1(a) and consists of three pairs of a BERT model and a subsequent softmax layer, where each pair computes the probability $P^{lbl}_{ann}(y^{ann}|\mathbf{x})$ of labels y^{ann} of each annotator $ann \in \{A, B, C\}$, in the same manner as Equation (2) in Table 1. Probability $P(y^{\mathrm{MV}}|\mathbf{x})$ of final label y^{MV} is the average of $P^{lbl}_{ann}(y^{ann}|\mathbf{x})$.

ProposedSingle: This method uses multi-task learning in which each task corresponds to predicting labels given by one of the three annotators. The architecture of this method is shown in Figure 1(b) and consists of a single BERT model with three softmax layers, where the output of each softmax layer corresponds to an annotator's label. $P(y^{\mathrm{MV}}|\mathbf{x})$

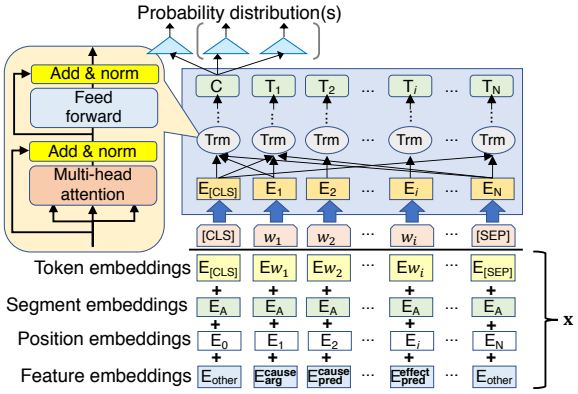


Figure 2: Proposed BERT architecture and its input $\mathbf{x}$

is computed in the same way as in **Proposed-Multi**.

ProposedRU: This method uses the **Proposed-Single** architecture, which consists of a BERT model (BERT_{lbl} in Figure 1(c)) and three softmax layers, to compute probability $P^{lbl}_{ann}(y^{ann}|\mathbf{x})$ for annotator ann. We also add another BERT model (BERT_{ru}) and its subsequent softmax layer to the architecture to assign a lower weight to the predictions of an annotator who is likely to disagree with the majority vote. The entire computation is done using the equations in Table 1. Since each causality candidate is independently labeled by three annotators, at most one annotator disagrees with the majority vote label. To identify that annotator, BERT_{ru}, along with the softmax layer, estimates the probability $P^{ru}(ann|\mathbf{x})$ that annotator ann disagrees with the majority vote. We call this probability the *relative unreliability* of ann (Equation (3)). Instead of averaging $P^{lbl}_{ann}(y^{ann}|\mathbf{x})$ as in **ProposedSingle**, **ProposedRU** uses the weighted sum of $P^{lbl}_{ann}(y^{ann}|\mathbf{x})$ to predict the final label (Equation (1)). Weight w_{ann} is computed from $P^{ru}(ann|\mathbf{x})$ to consider the relative unreliability of ann (Equation (4))[1].

Representation $\mathbf{x}$ of the input causality candidate is computed from the entire sentence (e.g., "He smoked cigarettes and died of lung cancer caused by them") that contains a pair of cause phrase ("smoke cigarettes") and effect phrase ("died of lung cancer"). $\mathbf{x}$ consists of the token, position, and segment embeddings of each word (Devlin et al., 2019). We distinguish the words in the cause and effect candidates from the other words by giving them *feature embeddings*; the cause argument, the cause predicate, the effect argument, the effect predicate, and the other words are all given a different feature embedding vector that is randomly initialized (see Figure 2).

$$L = \alpha L^{\text{MV}} + \beta \sum_{ann} L^{lbl}_{ann} + \gamma L^{ru} \quad (5)$$

$$L^{\text{MV}} = -\log P(y^{\text{MV}}|\mathbf{x}) \quad (6)$$

$$L^{lbl}_{ann} = -\log P^{lbl}_{ann}(y^{ann}|\mathbf{x}) \quad (7)$$

$$L^{ru} = \begin{cases} 0 & \text{if } y^A = y^B = y^C \\ -\log P^{ru}(ann|\mathbf{x}) & \text{otherwise} \end{cases} \quad (8)$$

Table 2: Loss functions used in ProposedRU

Note that we followed the fine-tuning scheme presented in Devlin et al. (2019) in the training of **ProposedMulti** and **ProposedSingle**. For training **ProposedRU**, we designed a special loss function L in Equation (5) in Table 2 as the weighted sum of L^{MV}, L^{lbl}_{ann}, and L^{ru}, each of which is a loss function for $P(y^{\text{MV}}|\mathbf{x})$, $P^{lbl}_{ann}(y^{ann}|\mathbf{x})$, and $P^{ru}(ann|\mathbf{x})$, respectively. Here, α, β, and γ are hyper-parameters ($\alpha + \beta + \gamma = 1.0$).

The BERT model in our methods is pre-trained from scratch using causality-rich texts to investigate whether such BERT models can learn background knowledge during their pre-training. We used 19,567,386 sentences from 2,799,079 passages extracted from four billion web pages, where each passage consists of seven sentences and includes at least one event causality detected by a CRF-based causality recognizer[2] (Oh et al., 2013) (3,046,619 event causalities were detected in the passages).

We also introduce **ProposedRU+BK**, which integrates the background knowledge used in the previous work into **ProposedRU**. As input, **ProposedRU+BK** is given a *pseudo sentence*, which is the concatenation of the original input sentence and Kruengkrai et al. (2017)'s text fragments embodying background knowledge[3] along with a

[1] Note that this method is applicable only for the problem setting in which the final binary-class label is determined by using three annotators' labels. Some extension is needed for computing relative unreliability $P^{ru}(ann|\mathbf{x})$ for an arbitrary number of annotators by considering the probability that each annotator agrees with the majority.

[2] The recognizer was applied only to passages that were extracted by specifically focusing on clue terms such as "because." Event causalities recognized this way are a different type than those under focus in this work.

[3] We used all three types of text fragments from Krueng-

separator to compute input representation $\mathbf{x}$.

Here, we hypothesized that BERT models could learn some sort of background knowledge by using causality-rich texts in pre-training. To investigate whether this is true, we pre-trained several BERT models, each using either causality-rich texts or general texts, and then fine-tuned each of the models using the same architecture as used in **ProposedRU**. We introduce **ProposedRU$_{\text{wiki}}$** and **ProposedRU$_{\text{web}}$**, in which the pre-training is performed from scratch using a single corpus, that is, either Wikipedia articles (19,567,381 sentences retrieved in August 2018) or randomly sampled web texts (19,567,396 sentences in 1,990,472 web pages randomly sampled from the four billion web pages), respectively. The fine-tuning is then done as in **ProposedRU**.

In contrast to the above single-step pre-training, we also attempted to start from the pre-trained model for **ProposedRU$_{\text{web}}$**, which was pre-trained using general web text, and then additionally pre-train it with causality-rich texts to make the model suitable for causality-event recognition[4]. More precisely, to focus on the event causality part described in each passage, we extracted the pairs of cause and effect parts detected by Oh et al. (2013)'s CRF-based causality recognizer and used them as a causality-rich corpus while assuming that the cause and effect are sentence pairs for the next-sentence prediction task (9,783,691 pairs or 19,567,382 sentences). This model is called **ProposedRU$_{\text{web+pair}}$** hereafter.

For comparison, we also introduce **ProposedRU$_{\text{web+web}}$**, which uses another set of web sentences randomly sampled from four billion web pages, as a corpus of the general but not causality-rich texts. The size is the same as that of the cause-effect pairs used for **ProposedRU$_{\text{web+pair}}$**.

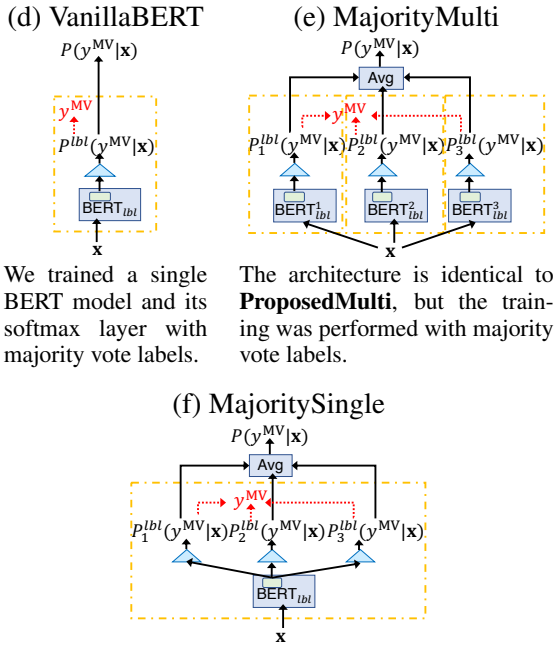
3 Experiments

3.1 Settings

We used the datasets for event-causality recognition in Japanese of Hashimoto et al. (2014). They regard causality candidate $A \rightarrow B$ proper if and only if "if A happens, the probability of B increases." To annotate this dataset, three annotators used their own judgment independently, and each annotator's individual labels are included in the datasets. Table 3 shows their statistics[5].

Data	#Instances	#True causalities
Training	107,068	8,986
Development	23,602	3,759
Test	23,650	3,647

Table 3: Statistics of datasets



We trained a single BERT model and its softmax layer with majority vote labels.

The architecture is identical to **ProposedMulti**, but the training was performed with majority vote labels.

The architecture is identical to **ProposedSingle**, but the training was performed with majority vote labels.

Figure 3: Baseline architectures

As baselines, we used the state-of-the-art method of Kruengkrai et al. (2017), **MCNN**, which uses a CNN and exploits text fragments retrieved from four billion web pages as background knowledge, and three BERT-based methods, **VanillaBERT**, **MajorityMulti**, and **MajoritySingle**, which use only majority vote labels (Figure 3). These BERT-based baselines used the same pre-trained BERT model (pre-trained with causality-rich texts) as in our proposed methods.

As the parameter settings of the pre-training for all of the BERT models, we followed the BERT$_{\text{BASE}}$ settings[6] (Devlin et al., 2019) except for a batch size of 50. For **ProposedRU$_{\text{web+pair}}$** and **ProposedRU$_{\text{web+web}}$**, we additionally pre-

krai et al. (2017) (i.e., short binary patterns, why-QA system's answers, and sentences with clue terms). Every text fragment was extracted from four billion Japanese web pages.

[4] This approach is recommended by Google's BERT implementation (https://github.com/google-research/bert) for computational efficiency.

[5] The number of training instances in Hashimoto et al. (2014)'s dataset was 112,098, but we excluded the duplicates in it and used the resulting 107,068 instances for training each method.

[6] 12-layers, 768 hidden states, 12 heads, and training for one million steps with a warmup rate of 1% using an Adam optimizer with a learning rate of 1e-4.

Model	R	P	F	Avg.P
MCNN	40.2	**61.1**	48.5	55.1
VanillaBERT	63.6	49.7	55.8*	54.2
MajorityMulti	61.5	51.8	56.2*	55.7
MajoritySingle	48.3	56.8	52.2*	54.4
ProposedMulti	63.9	51.3	56.9*	56.7
ProposeSingle	62.8	52.7	57.3*†	57.1
ProposedRU	**64.0**	52.0	**57.4***	**57.4**

'*' stands for significant improvement over MCNN and '†' means that over MajorityMulti (McNemar test, $p = 0.05$).

Table 4: Results of event-causality recognition

Model	R	P	F	Avg.P
ProposedRU$_{\text{wiki}}$	57.4	49.6	53.2$\star$	53.3
ProposedRU$_{\text{web}}$	59.0	50.9	54.6$\star$	54.5
ProposedRU	64.0	52.0	57.4	57.4
ProposedRU$_{\text{web+web}}$	62.5	49.0	54.9$\star$	54.8
ProposedRU$_{\text{web+pair}}$	64.3	48.2	55.1$\star$	55.3
ProposedRU+BK	**67.4**	**52.3**	**58.9**	**59.9**

'$\star$' stands for significant difference from ProposedRU.

Table 5: Results of ProposedRU and its variants

trained the models for 0.3 million steps with a learning rate of 2e-5. As for the fine-tuning, we chose the combination of epochs {1, 2, 3} and learning rates {5e-6, 1e-5, 2e-5, 3e-5, 4e-5, 5e-5} that achieved the best average precision on 50% of the development dataset, as done in Kruengkrai et al. (2017). For **ProposedRU** and its variants, we additionally optimized the α, β, and γ hyper-parameters in the same way.

3.2 Results

The recall, precision, F1-score, and average precision of each method on the majority vote labels of the test dataset are presented in Table 4. **ProposedRU** achieved the best F1-score and average precision. All of the proposed methods (**ProposedMulti**, **ProposedSingle**, and **ProposedRU**) outperformed all of the baseline methods (**MCNN**, **VanillaBERT**, **MajorityMulti**, and **MajoritySingle**) in average precision. This suggests that using each annotator's labels produced a positive effect. The F1-scores of **ProposedRU** and **ProposedSingle** were not significantly different (McNemar test, $p = 0.05$).

Table 5 shows the results of our investigation into the pre-training of some sort of *background knowledge*. Among the methods to be compared in Table 5, the first three methods (i.e., **ProposedRU$_{\text{wiki}}$**, **ProposedRU$_{\text{web}}$**, and **ProposedRU**) utilized a single-step pre-training from scratch, whereas the next two methods (**ProposedRU$_{\text{web+web}}$** and **ProposedRU$_{\text{web+pair}}$**) performed additional pre-training before the fine-tuning. The results show that pre-training using causality-rich texts contributes to further performance improvement than that using general texts, such as Wikipedia articles and random web texts (see **ProposedRU** vs. **ProposedRU$_{\text{wiki}}$**, **ProposedRU** vs. **ProposedRU$_{\text{web}}$**, and **ProposedRU$_{\text{web+pair}}$** vs. **ProposedRU$_{\text{web+web}}$**). In short, we can say that BERT might learn some sort of background knowledge from causality-rich texts. An interesting point is that, although the amount of texts used for **ProposedRU** is almost the same as that of the second-step causality-rich pre-training for **ProposedRU$_{\text{web+pair}}$**, the performance differs considerably (about 2% difference). This suggests that the design of the pre-training steps is not straightforward and thus merits further research.

We further evaluated **ProposedRU+BK**, in which text fragments embodying background knowledge are concatenated to the input sentence as explained in Section 2. Table 5 shows that **ProposedRU+BK** improved the average precision over **ProposedRU** by about 2.5% (i.e., **ProposedRU+BK** significantly outperformed the state-of-the-art method, **MCNN**, by about 5%), suggesting that background knowledge in the form of text fragments is still useful, at least in our current experimental setting. However, the usefulness might be lost when a model is appropriately pre-trained with a larger amount of texts that covers even more background knowledge.

4 Conclusion

This paper proposed BERT-based methods for recognizing event causality that exploit each annotator's independent judgments. By using each annotator's judgments, we showed that even a simple multi-task learning approach or an ensemble method improved performance in our experiments. Our best-performing method significantly outperformed the state-of-the-art method by about 5% in average precision. We also confirmed that the pre-trained BERT model learns some sort of background knowledge for event causality from causality-rich texts, although text fragments embodying background knowledge remain useful, at least in our current setting. As future work, we are examining the possibility of using an adversarial learning framework (Goodfellow et al., 2014), which was recently used in the why-QA task (Oh et al., 2019) for causality recognition.

References

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.

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.

Paul Felt, Eric Ringger, and Kevin Seppi. 2016. Semantic annotation aggregation with conditional crowdsourcing models and word embeddings. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 1787–1796.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 2672–2680.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 619–630.

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*, pages 987–997.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3466–3473.

Jiyi Li, Yukino Baba, and Hisashi Kashima. 2017. Hyper questions: Unsupervised targeting of a few experts in crowdsourcing. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1069–1078.

Jong-Hoon Oh, Kazuma Kadowaki, Julien Kloetzer, Ryu Iida, and Kentaro Torisawa. 2019. Open-domain why-question answering with adversarial learning to encode answer texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4227–4237.

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*, pages 1733–1743.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. *2010 IEEE Fourth International Conference on Semantic Computing*, pages 361–368.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Kentaro Torisawa. 2006. Acquiring inference rules with temporal constraints by using Japanese coordinated sentences and noun-verb co-occurrences. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 57–64.

Zhipeng Xie and Feiteng Mu. 2019. Distributed representation of words in cause and effect spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7330–7337.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229.

Dengyong Zhou, Sumit Basu, Yi Mao, and John C. Platt. 2012. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems 25*, pages 2195–2203.