NoisyAnnot@ Causal News Corpus 2022: Causality Detection using Multiple Annotation Decisions

Quynh Anh Nguyen[1, 2] **Arka Mitra**[2]
[1] University of Milan [2] ETH Zürich
quynguyen@ethz.ch, amitra@ethz.ch

Abstract

The paper describes the work that has been submitted to the 5^{th} workshop on Challenges and Applications of Automated Extraction of socio-political events from text (CASE 2022). The work is associated with Subtask 1 of Shared Task 3 that aims to detect causality in protest news corpus. The authors used different large language models with customized cross-entropy loss functions that exploit annotation information. The experiments showed that *bert-based-uncased* with refined cross-entropy outperformed the others, achieving a F1 score of 0.8501 on the Causal News Corpus dataset.

1 Introduction

A causal relationship in a sentence implies an underlying semantic dependency between the two main clauses. The clauses in these sentences are generally connected by markers which can have different parts of tags in the sentence. Moreover, the markers can be either implicit or explicit and for these reasons, one cannot rely on regex or dictionary-based systems. Thus, there is a need to investigate the context of the sentences. For the given task, we exploited different large language models that provide a contextual representation of sentences to tackle causality detection.

Shared task 3 in CASE-2022 (Tan et al., 2022a) aims for causality detection in news corpus, which can be structured as a text classification problem with binary labels. Pre-trained transformer-based models (Vaswani et al., 2017) have shown success on tackling a wide range of NLP tasks including text generation, text classification, etc. The authors look into inter-annotation agreements and number of experts and how they can be included in the loss to improve the performance of the pre-trained models.

The main contributions of the paper are as follows:

1. Extensive experimentation with different large language models.
2. Incorporation of additional annotation information, i.e inter-annotation agreement and the number of annotators, to the loss.

The remaining paper is formulated as follows: Section 2 reviews the related work, section 3 describes the dataset on which the work has been done, section 4 discusses the methodology used in the paper, the following section discusses the results and provides an ablation of the various loss functions introduced and finally, section 6 concludes the paper and suggests future works.

2 Related Work

Multiple annotations on a single sample reduce the chances of the labelling to be incorrect or bias being incorporated into the dataset (Snow et al., 2008). Including multiple annotators also leads to disagreement among the labels that have been provided by them. The final or gold annotation is then usually determined by majority voting (Sabou et al., 2014) or by using the label of an "expert" (Waseem and Hovy, 2016). There are also different methodologies which do not use majority voting to select the "ground truth".

Expectation Maximization algorithm has been used to account for the annotator error (Dawid and Skene, 1979). Entropy metrics have been developed to identify the performance of the annotators(Waterhouse, 2012; Hovy et al., 2013; Gordon et al., 2021). Multi-task learning is also used to deal with disagreement in the labels (Fornaciari et al., 2021; Liu et al., 2019; Cohn and Specia, 2013; Davani et al., 2022). There are methods which include the annotation disagreement into the loss function for part of speech tagging (Plank et al., 2014; Prabhakaran et al., 2012) on SVMs and perceptron model. The present work considers the inter-annotator agreement as well as the number

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 79 - 84
December 7-8, 2022 ©2022 Association for Computational Linguistics

of annotators into the loss function for any model. The work also compares the performance when the annotators who disagree with the majority voting has been ignored.

3 Dataset

The Causal News Corpus dataset (Tan et al., 2022b) consists of 3,559 event sentences extracted from protest event news. Each sample in the dataset contains the text, the corresponding label, the number of experts who annotated the label and the degree of agreement among the experts. Figure 1 shows a sample from the provided training set. The training data is fairly balanced, containing 1603 sentences with a causal structure and 1322 sentences without a causal structure. Also, the number of causal and non-causal sentences in the validation set does not differ significantly. Finally, 311 news articles have been used as test set for evaluation.


```
{'index': 'train_01_10',
 'text': "The farmworkers ' strike
resumed on Tuesday when their demands
were not met .",
 'label': 1,
 'agreement': 1.0,
 'num_votes': 3,
 'sample_set': 'train_01'}
```


Figure 1: A datapoint from the provided training data.

Besides the binary labels, the Causal News Corpus dataset also provides additional information regarding the number of experts who labeled the sentence and the percentage of agreement between them. Figure 1 shows that the number of experts who annotated the text *"The farmworkers'strike resumed on Tuesday when their demands were not met."* is 3 (`num_votes` = 3). Also, all of the experts labeled the sentence to be causal so the agreement is 1.0 (100% agreement) and the label is 1. In case only one of three experts assigned label 1 to the previous text, the three predictors `num_votes`, `agreement`, `label` would now become 3, $\frac{2}{3}$, 0 respectively. In this paper, the authors exploit this information to give the model more prior and thus potentially improve the model's performance, which has been described in more detail in section 4.

4 Methodology

The section discusses the pipeline, the different types of loss functions that were implemented, and the experimental details that have been used in the third shared task for CASE 2022 (Tan et al., 2022a).

4.1 Pipeline

The authors finetuned large language models with different loss functions to tackle Subtask 1 in Shared Task 3 of CASE@EMNLP-2022, causality detection in a given sentence. The problem can be reformulated as a binary classification where the model predicts whether the sentence is causal or not. Since contextual awareness plays an essential role in handling this specific task, the authors used several transformer-based models, namely, BERT (Devlin et al., 2019), FinBERT (Liu et al., 2020), XLNET (Yang et al., 2019) and RoBERTa (Zhuang et al., 2021).

The given sentence is first tokenized by a tokenizer from the corresponding pretrained model architecture provided by HuggingFace (Wolf et al., 2020). The vector output from the tokenization stage is then fed as input to the model. The most informative token is the classification token ([CLS]), which is a special token that can be used as a sentence representation. The [CLS] token is then passed through a feed-forward network to generate logits. The softmax over the logits gives us the probability of whether the sentence is causal or not. For each model, the authors experimented with cross-entropy loss and proposed two loss functions described in detail in subsection 4.2.

4.2 Loss Functions

Cross Entropy Loss The loss of the classification task can be represented by a simple cross-entropy loss, as shown in Equation 1:

$$L = \frac{1}{M}\sum_{i=1}^{M}(-y_i^{true}log(y_i^{pred}) - (1-y_i^{true})log(1-y_i^{pred})) \quad (1)$$

where y_i^{true} and y_i^{pred} denote the true label and the predicted label for the i^{th} input in a batch of M sentences.

Noisy Cross Entropy Loss The dataset not only provides the standard information about `{text, label}`, but also contains the information about the number of experts who annotated the sentence's label, and proportion of agreement between them. The authors have considered the annotation by each of the experts to be the true label for the sentence. For a sentence with n expert annotations (`num_votes` $= n$) and r percent of agreement (`agreement` $= r$), the loss for each sentence can be written as shown in Equation 2.

$$L = \begin{cases} (-rlog(y^{pred}) \\ -(1-r)log(1-y^{pred})), & \text{if } y^{true} = 1, \\ (-(1-r)log(y^{pred}) \\ -rlog(1-y^{pred})), & \text{if } y^{true} = 0. \end{cases} \tag{2}$$

The equations can be combined and the loss for a batch of M sentences can be rewritten as:

$$\begin{aligned} L = & \frac{1}{\sum_{i=1}^{M} n_i} \sum_{i=1}^{M} (-y_i^{true} n_i (r_i log(y_i^{pred}) \\ & + (1-r_i) log(1-y_i^{pred})) \\ & - (1-y_i^{true}) n_i (r_i log(1-y_i^{pred}) \\ & + (1-r_i) log(y_i^{pred}))) \end{aligned} \tag{3}$$

.

The different annotations from all the experts has been considered, adding more information to the model. Equation 3 takes the n votes from the different experts into account, out of which $n \times r$ times it is assigned the correct label, and the incorrect label has been used the other $n \times (1-r)$ times. If the labels from the different experts are taken directly, there will be conflicts in the labels when the experts disagree. Considering the loss for one sentence when the true label is 1, the derivative of the loss is shown in Equation 4. Figure 2 shows that the loss is minimized when y^{pred} is equal to r and its minima shifts from 1 to 0 as the level of agreement decreases when the true label is 1. A similar profile is obtained when the true label is considered to be 0. The formulation pushes the solution to a distribution where the ideal output is not a one-hot encoding, which is similar to the label smoothing method. Label smoothing was initially proposed by Szegedy et al. (Szegedy et al., 2016) to improve the performance of the Inception architecture on the ImageNet dataset (Deng et al., 2009). In label-smoothing, the ground truth sent to the model is not encoded as a one-hot representation. Since there are conflicts in the annotations and the loss considers all of the noisy data, it has been referred as noisy cross-entropy loss.

$$\frac{\partial L}{\partial y^{pred}} = \frac{y^{pred} - r}{y^{pred}(1-y^{pred})} \tag{4}$$

Refined Cross Entropy Loss The ideal output of the model should be close to the ground truth label. Thus, a modification to loss function should be done to improve the performance. The error occurs when the annotators who have not agreed for a particular label have also been taken into consideration. The number of experts who provided the correct label can also be an important signal to the model. If a sentence has been given a label by a more significant number of experts, the model should be penalized more if the sentence is misclassified. The new loss, over a batch of M sentences, can thus be written as :

$$\begin{aligned} L = & \frac{1}{\sum_{i=1}^{M} n_i r_i} \sum_{i=1}^{M} (-y_i^{true} n_i r_i log(y_i^{pred}) \\ & - (1-y_i^{true}) n_i r_i log(1-y_i^{pred})) \end{aligned} \tag{5}$$

.

The number of causal and non-causal sentences is almost the same and there is no significant class imbalance. The authors have thus not considered weight penalization to the class with the higher number of samples.

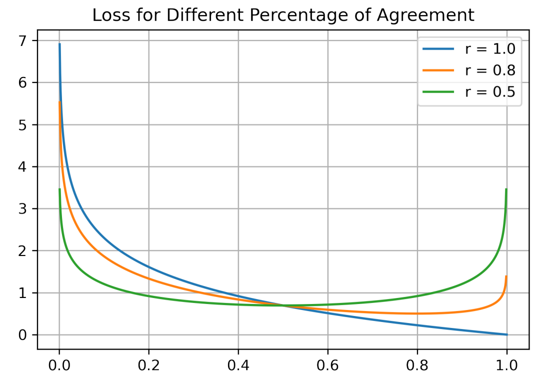


Figure 2: Loss for noisy cross-entropy

4.3 Experimental Details

The experiments have been performed in PyTorch (Paszke et al., 2019) and the authors used the HuggingFace (Wolf et al., 2020) library to generate the pipeline for the different experiments. Each model has been trained for 10 epochs with a learning rate of 5×10^{-5} and a seed of 42 for reproducibility. Various models have been considered and trained with the same set of hyperparameters. The code is made publicly available on Github [1].

[1] `https://github.com/jyanqa/case-2022-causual-event`

Model name	**Cross Entropy**	**Noisy Cross Entropy**	**Refined Cross Entropy**
bert-based-cased (Devlin et al., 2019)	**0.8251**	0.8225	0.8235
bert-base-uncased (Devlin et al., 2019)	0.8283	0.8313	**0.8501**
bert-large-cased (Devlin et al., 2019)	0.7105	**0.7549**	0.7105
xlnet-based-cased (Yang et al., 2019)	0.7953	**0.8216**	0.8199
roberta-base (Zhuang et al., 2021)	0.8279	0.8279	**0.8280**

Table 1: Evaluation of models on different loss functions. The best F1 score of each model is marked in bold.

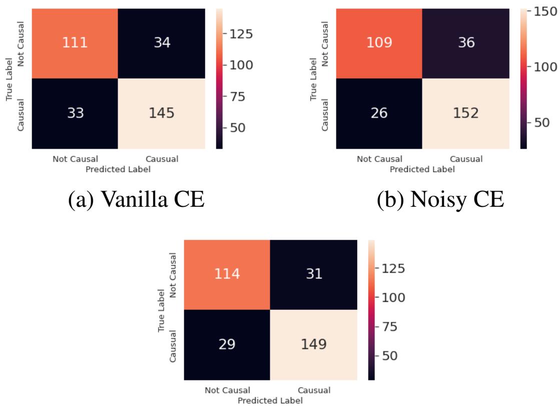
5 Results and Discussion

In this section, the results of the different models and the different losses are discussed.

Table 1 shows the evaluation of the different models on the validation set. Performances of four in five models, excepting the *bert-base-uncased* case, are enhanced by leveraging the modified cross-entropy loss. In fact, the F1 scores of four models are significantly increasing when we replaced vanilla cross-entropy loss with noisy cross-entropy loss and refined cross-entropy loss. Specifically, model fine-tuned from *bert-base-uncased* investigating Refined cross-entropy loss function yields the best performance in all experimented models with F1 score of 0.8501. On the other hand, *bert-base-cased* is the only pretrained model that does not benefit from customized cross-entropy losses. Adapting vanilla cross-entropy function on *bert-base-cased* model results in its best F1 scores of 0.8251.

The models with noisy and refined cross-entropy loss utilizes the annotated information and thus performs better. The noisy cross-entropy loss is similar to restricting the highest probability output that a model can predict. However, in almost all cases, the degree of agreement was either 1 or $\frac{2}{3}$. In general, the smooth labelling has a value in the range of 0.9 to 1. Different contradicting annotations of labels might make the model face difficulties in learning and yielding an accurate prediction for each sentence. The refined cross-entropy solely considers the labels that do not contradict each other, thus it performs the best.

Moreover, the experiments show that *roberta-based* models achieve lower performance compared to BERT-based models, especially *bert-base-uncased* models. The model pretrained on *bert-large-cased* has been fine-tuned for only one epoch due to computation limitations. Their F1 scores are worse than those of *bert-base-cased* and *bert-base-uncased* models. *bert-base* models result in better performance, as compared to models fine-tuned on *roberta-base*. The reason could be that RoBERTa-based models had not been trained on next sentence prediction (NSP) while BERT-based models were. Causality detection can benefit from NSP. A sentence can be considered to be two relevant clauses that are joined by a causal effect. Thus, knowing if the clauses are relevant or not benefits the task of causality detection.



(a) Vanilla CE (b) Noisy CE

(c) Refined CE

Figure 3: Confusion matrix for the different losses

Figure 3 shows the confusion matrix resulting from *bert-base-uncased* models which result the best F1 scores in all implemented models. Models are generally good at predicting non-causal sentences regardless of the loss function used. In fact, true negatives and true positives are always the highest measures compared to the others. On the other hand, there is a clear trend in the number of true positives when we shift the loss function from vanilla to noisy and refined cross-entropy. In particular, the model yields 145 true positives and is improved to 152 and 149 true positives when we replaced vanilla cross-entropy loss with noisy and refined cross-entropy loss function.

6 Conclusion

This paper presents our work on detecting causal effect relationships in news corpus by fine-tuning Transformers-based models and adapting multiple loss functions. The experiments showed that considering annotation information using customized loss functions significantly improved the model performance in four out of five experimented models. Besides, the experiments show that BERT outperformed RoBERTa, which can be attributed to the fact that RoBERTa is not trained on NSP. Last but not least, the *bert-base-uncased* obtained the best performance amongst all 15 models with an F1-score of 0.8501 in validation set and 84.930 in the test set using the refined cross-entropy loss that takes account of the annotation information presented in the dataset.

The authors plan to look into exploiting the uncertainty of the annotator's information and parameterizing the loss function to further enhance the model's performance.

References

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *ACL*.

Aida Mostafazadeh Davani, Mark D'iaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

A. Philip Dawid and Allan Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*, 28:20–28.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *NAACL*.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori B. Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with mace. In *NAACL*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *ACL*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *IJCAI*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *EACL*.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona T. Diab, B. Dorr, Lori S. Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *ExProM@ACL*.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*.

Rion Snow, Brendan T. O'Connor, Dan Jurafsky, and A. Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

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

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

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL*.

Tamsyn P. Waterhouse. 2012. Pay by the bit: an information-theoretic metric for collective human judgment. *Proceedings of the 2013 conference on Computer supported cooperative work*.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.