The Causal News Corpus: Annotating Causal Relations in Event Sentences from News

Connect with us!

https://github.com/tanfiona/CausalNewsCorpus

tan.f@u.nus.edu

Fiona Anting Tan[1], Ali Hürriyetoğlu[2], Tommaso Caselli[3], Nelleke Oostdijk[4], Tadashi Nomoto[5], Hansi Hettiarachchi[6], Iqra Ameer[7], Onur Uca[8], Farhana Ferdousi Liza[9], Tiancheng Hu[10]

[1] Institute of Data Science, National University of Singapore, Singapore, [2] Koc University, Turkey, [3] Rijksuniversiteit Groningen, Netherlands, [4] Radboud University, Netherlands, [5] National Institute of Japanese Literature, Japan, [6] Birmingham City University, United Kingdom, [7] Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico, [8] Department of Sociology, Mersin University, Turkey, [9] University of East Anglia, United Kingdom,[10] ETH Zürich, Switzerland

Introduction

Causality is a core cognitive concept and appears in many NLP works on inference and understanding (Jo et al., 2021; Dunietz et al., 2020; Feder et al., 2021). A causal relation is a semantic relationship between two arguments known as cause and effect, in which the occurrence of one (Cause argument) causes the occurrence of the other (Effect argument) (Barik et al., 2016). Causality can be expressed in various ways: explicitly, implicitly, or through alternative lexicalizations.

Causal

The treating doctors said Sangram lost around 5 kg due to the hunger strike .

The bombing created panic among villagers .

Dissatisfied with the package , workers staged an all-night sit-in .

Non-causal

Thus we too joined the sloganeering .

The alliance claimed 4,000 took part last year .

Figure 1: Annotated examples from Causal News Corpus. Causes are in pink, Effects in green and Signals in yellow. Note that both Cause and Effect spans must be present within one and the same sentence for us to mark it as *Causal*.

Motivation

The extraction of causality from text is challenging because semantic understanding of the context and world knowledge is needed.

Existing corpora on event causality (E.g. CausalTimeBank (CTB) (Mirza et al., 2014), CaTeRS (Mostafazadeh et al., 2016), EventStoryLine (Caselli and Vossen, 2017)) are limited in size.

There is also a discrepancy between such event causality corpora and other causality corpora. Penn Discourse Treebank (PDTB) (Prasad et al., 2008; Webber et al., 2019; Prasad et al., 2006) is a corpus that annotates semantic relations (including causal relations) between clauses, expressed in all constructions. This corpus is large and potentially useful for training an accurate event sentence classifier. Therefore, we believe that it will be beneficial to align the annotation guidelines of these two corpora types. Currently, they differ in what they regard as arguments.

Many corpora only focus on explicit relations (Girju and Moldovan, 2002; Dunietz et al., 2017), and likewise for CTB. Implicit relations are more common but more challenging to identify (Hidey and McKeown, 2016).

CNC builds on the datasets featured in a series of workshops aimed at mining socio-political events from news articles: AESPEN 2020 and CASE 2021.

References:

- Barik, B., Marsi, E., and Öztürk, P. (2016). Event causality extraction from natural science literature. Res. Comput. Sci., 117:97–107.
- Caselli, T. and Vossen, P. (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, August. Association for Computational Linguistics.
- Dunietz, J., Levin, L., and Carbonell, J. (2017). The BECauSE corpus 2.0: Annotating causality and overlapping relations. In Proceedings of the 11th Linguistic Annotation Workshop, pages 95–104, Valencia, Spain, April. Association for Computational Linguistics.
- Dunietz, J., Burnham, G., Bharadwaj, A., Rambow, O., Chu-Carroll, J., and Ferrucci, D. (2020). To test machine comprehension, start by defining comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7839–7859, Online, July. Association for Computational Linguistics.
- Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. (2021). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. arXiv preprint arXiv:2109.00725.
- Girju, R. and Moldovan, D. I. (2002). Text mining for causal relations. In Susan M. Haller et al., editors, Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, May 14-16, 2002, Pensacola Beach, Florida, USA, pages 360–364. AAAI Press.
- Grivaz, C. (2010). Human judgements on causation in French texts. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, May. European Language Resources Association (ELRA).
- Hidey, C. and McKeown, K. (2016). Identifying causal relations using parallel Wikipedia articles. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1424–1433, Berlin, Germany, August. Association for Computational Linguistics.
- Jo, Y., Bang, S., Reed, C., and Hovy, E. H. (2021).Classifying argumentative relations using logical mechanisms and argumentation schemes. Trans. Assoc.Comput. Linguistics, 9:721–739.
- Mirza, P., Sprugnoli, R., Tonelli, S., and Speranza, M. (2014). Annotating causality in the TempEval-3 corpus. In Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), pages 10–19, Gothenburg, Sweden, April. Association for Computational Linguistics.
- Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J., and Vanderwende, L. (2016). CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In Proceedings of the Fourth Workshop on Events, pages 51–61, San Diego, California, June. Association for Computational Linguistics.
- Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., and Webber, B. L. (2006). The penn discourse treebank 1.0 annotation manual. IRCS Technical Reports Series.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., andWebber, B. L. (2008). The penn discourse treebank 2.0. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco. European Language Resources Association.
- Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The penn discourse treebank 3.0 annotation manual. Philadelphia, University of Pennsylvania.

Compilation & Annotation

I. Guidelines

We labeled sentences to be *Causal* based on adaptations of the definition for CONTINGENCY from PDTB-3. We also utilized the five tests for causality based on the work by Grivaz (2010) and Dunietz et al. (2017).

Sentence	Causality Tests					Label
	Why?	Temporal Order	Counterfact.	Ontological Asymmetry	Linguistic	
The protests spread to 15 other towns and resulted in two deaths and the destruction of property .	✓	✓	✓	✓	✓	*Causal*
Chale was allegedly chased by a group of about 30 people and was hacked to death with pangas, axes and spears .	✗	✓	✗	✓	✗	*Not Causal*
The strike will continue till our demands are conceded .	✓	✓	✓	✓	✓	*Causal (Neg. Cond.)*

Table 1: Examples illustrating the applications of the Tests for Causality. Cause in pink, Effect in green; potential Cause in gray. Signals are not marked.

II. Workflow

Five annotators and one curator were involved. Iterative feedback and guideline refinements were needed to improve agreement scores. The overall Krippendorff's Alpha score was 34.99%.

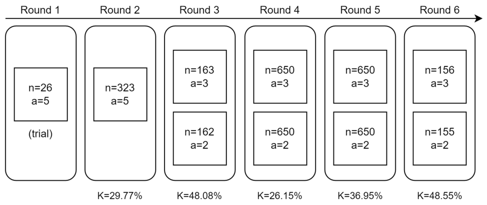

Figure 2: Summary of annotation workflow. Round 1 was a trial training round where annotations were discarded. All subsequent rounds were for the training set except Round 6 which was for the test set.

III. Final Dataset

The Causal News Corpus (CNC) contains 3,559 annotated sentences, with 1,957 marked as *Causal* and 1,602 marked as *Non-causal*.

Conclusion

CNC's annotation guidelines covers a wider array of causal linguistic constructions than previous works. We demonstrated transferability between CNC and existing datasets that include causal relations. CNC, which has been annotated by experts, is a valuable resource for researchers. We are also organizing a shared task using CNC,

Experiments

I. Testing on CNC

When training and testing on CNC, we achieved a reasonable F1 score of 81.20%, exceeding the dummy baselines scores, demonstrating that our annotations are internally consistent and reliable. When training on external corpora and testing on CNC, we achieved up to ~64% F1 without additional fine-tuning. This demonstrates the transferability of existing causality corpora on CNC. Since external corpora and CNC differs slightly in annotation guidelines, some performance differences are expected.

#	Training Set	F1	P	R	Acc	MCC
1	All *Causal*	72.28	56.59	**100.00**	56.59	0.00
2	Random	55.72	56.61	54.92	50.66	0.00
3	CNC Training	**81.20**	78.01	84.66	**77.81**	**54.52**
4	PDTB-3	55.43	**81.32**	42.05	61.74	32.09
5	PDTB-3 Bal	64.45	77.60	55.11	65.59	34.75
6	CTB	27.36	80.56	16.48	50.48	17.49
7	CTB Bal	64.05	75.38	55.68	64.63	32.13

Table 2: Metrics from predictions on CNC Test Set.

II. Training/Testing across datasets

CNC is the most transferable between the event causality (i.e. CTB) and linguistic causality (i.e. PDTB) corpora studied.

Training Set	Test Set							
	F1				MCC			
	CNC	PDTB-3	CTB Bal	TRF↑	CNC	PDTB-3	CTB Bal	TRF↑
CNC	**83.46**	58.38	80.65	**74.16**	**61.71**	30.68	59.11	**50.50**
PDTB-3	56.45	**74.45**	60.79	63.90	35.86	**61.36**	32.49	43.24
CTB Bal	59.10	49.21	**83.41**	63.90	32.10	17.48	**65.01**	38.20

Table 3: Metrics from predictions using different train and test sets. Transferability Rate (TRF) indicates how well a model trained on a given corpus works for unseen, external datasets

III. CNC for Pre-training

Using a CNC-pretrained model (PTM) returns better performance than `bert-base-cased` PTM for out-of-domain datasets.

Dataset	PTM	F1	P	R	Acc	MCC
PDTB	`bert-base-cased`	74.45	**76.76**	72.31	82.60	61.36
	CNC-PTM	**75.19**	75.73	**74.69**	**82.71**	**61.95**
CTB Bal	`bert-base-cased`	83.41	77.25	**90.95**	81.91	65.01
	CNC-PTM	**84.68**	**80.14**	90.02	**83.80**	**68.30**

Table 4: Metrics from pre-trained model (PTM) experiments.

IV. Crowd-sourced Workers

A layman identifies causality poorly. Thus, CNC is a unique and valuable resource, requiring time and effort to create with experts.

	F1	P	R	Acc	MCC
All *Causal*	**66.00**	**50.00**	**100.00**	**48.40**	-3.89
Majority	61.97	47.83	88.00	46.00	-14.74
Each Vote	59.31	48.96	75.20	48.40	**-3.79**

Table 5: Metrics from crowd-sourced workers for a subset of 50 examples.