UniCausal: Unified Benchmark and Repository for Causal Text Mining – Supplementary Material

Fiona Anting Tan[1][0000−0002−2828−1831], Xinyu Zuo[2], and See-Kiong Ng[1]

[1] Institute of Data Science, National University of Singapore, Singapore
tan.f@u.nus.edu, seekiong@nus.edu.sg
[2] Tencent Technology, China xylonzuo@tencent.com
https://github.com/tanfiona/UniCausal

1 Data Processing

In this Section, we describe the processing steps done for each corpus. The codes are available under the '`processing`' folder in our repository.

AltLex [5] We used the development and gold datasets annotated by graduate students and crowd workers respectively. Although additional distantly labeled data are available in their repository, we wish to focus on a high-quality datasets and thus, incorporated only human-labelled datasets. Additionally, only the human-annotated datasets include labels with REASON and RESULT, which are needed to identify the Cause and Effect arguments. REASON and RESULT sentences were regarded as *Causal*, while sentences with the label NONE were treated as *Non-causal*. REASON examples were in the order of Effect-Signal-Cause, while RESULT had the format of Cause-Signal-Effect. Equipped with this template, we could convert the causal labels into Cause and Effect spans around the marked signal words. For train-test splits, the development set was used for testing, while the gold set was used for training. The AltLex corpus is suitable for all three tasks: For Span Detection, we had 415 examples; For sequence classification, we had 415 causal and 563 non-causal examples; For pair classification, we had access to 442 causal and 585 non-causal examples.

BECAUSE 2.0 [2] There are a few sentences where the Effect is annotated with no corresponding Cause[3] Such examples are treated as *Non-causal* in our consolidated corpus because we require both Cause and Effect spans to be present within each example for them to be considered *Causal*. On top of this treatment, BECAUSE 2.0 itself also contains some *Non-causal* examples. Additionally, we did not have access to the raw files of the NYT which requires subscription, and missed out on annotations for the 59 randomly selected articles. All in all, we obtained 965 *Causal* pairs and 280 *Non-causal* pairs. To fit our framework of creating a fixed subset for train-test split for easier benchmarking, we randomly set

[3] An example sentence from BECAUSE 2.0 where the Effect is annotated with no corresponding cause is: *"However, the lower prices these retail chains are now expected to bring should make it easier for managers to raise the necessary capital and pay back the resulting <effect>debt</effect>."*.

10% of the documents as test sets.[4] For PTB files, we followed the development and test splits recommended by PDTB V2.0 [10], mentioned later below.

CausalTimeBank (CTB) [6, 7] After filtering to examples with <= 3 sentence length, we ended up with 318 causal and 3,491 non-causal examples for pair classification, and 276 causal and 1,925 non-causal examples for sequence classification. To fit our framework of creating a fixed subset for train-test split for easier benchmarking, we randomly set 10% of the documents as test sets.[5] For files that tapped onto PTB, we followed the development and test splits recommended by PDTB V2.0 [10] which is described later below. Previous works [13, 7] evaluated models using 10-fold cross-validation or to use an external Temporal Eval-3 dataset [8]. Our method essentially only evaluates on one of the ten folds. Researchers can process the dataset to perform the other nine folds if needed. For Temporal Eval-3, we will work on including this dataset in the next release.

EventStoryLine (ESL) [1] We utilised the V1.0 expert annotations because it is the more updated version according to caselli-vossen-2017-event[6]. For events, we only consider markables tagged as actions or negative actions. All `PLOT_LINK` relations are treated as *Causal*.[7] To generate *Non-causal* examples, for every head event, we defined all tails unidentified by `PLOT_LINK` as *Non-causal*. Since ESL does not mark out the causal direction of their events, this dataset cannot be used for Pair Classification. ESL is only suitable for Sequence Classification, of which after filtering to keep only examples with <= 3 sentences, we obtained 2,232 examples. The last two topics were used as test set,[8] while the remaining 20 topics were used for training, as suggested by gao-etal-2019-modeling [3].

Penn Discourse Treebank V3.0 (PDTB) [12] The sense CONTINGENCY[9] was treated as a positive example of causal relations by previous researchers [9, 2]. Dunietz et al. [2] argued for the exclusion of evidentiary uses of causal language.[10] Thus, similar to the treatment by previous work [11], we treated the

[4] Test set: 'Article247_327.ann'

[5] Test set: 'ea980120.1830.0456.tml', 'APW19980227.0494.tml', 'PRI19980306.2000.1675.tml', 'APW19980213.1320.tml', 'APW19980501.0480.tml', 'PRI19980205.2000.1998.tml'

[6] Compared to V0.9, V1.0 includes additional data curation checks to ensure consistency of TimeX3 normalisation, `TLINK` annotation and event annotation, plus removes some wrong event tokens.

[7] We ignore markables tagged as location, time or time date. TLINK relations that have time related entities (E.g. *"May 02"* or *"Thursday"*) are not difficult to classify as *Non-causal*. Hence, we excluded them from our corpus.

[8] Test set: T37, T41

[9] CONTINGENCY: one argument provides the reason, explanation or justification for the situation described by the other [12]

[10] E.g. *"We went to war based on bad intelligence."* [2]

two sub-senses that were similar to evidential types, Belief and SpeechAct[11], as negative examples. For PDTB, we could retrace the Cause and Effect arguments in place based on the sub-sense. For five sub-senses,[12] the first argument is the Cause, while the second argument is the Effect. For four sub-senses,[13] the reverse. Examples with all other sub-senses were treated as *Non-causal*. The PDTB corpus is suitable for all three tasks. After filtering to examples with <= 3 sentences and <= 3 causal relations, we had 10,972 Span Detection, 41,669 Sequence Classification and 52,384 Pair Classification examples. For train-test split, we used the development and test set recommended by PDTB V2.0 [10],[14] while the remaining was used for training. Unless explicitly specified otherwise, we refer to the final test set as the combination of the development and test set.

SemEval 2010 Task 8 (SemEval) [4] We regarded the `Cause-Effect` relation as *Causal*, while all other relations were treated as *Non-causal*. Since the arguments annotated were short, SemEval is not compatible for Span Detection. Therefore, we only formatted SemEval for the Sequence Classification and Pair Classification task, of which, we obtained 10,690 and 10,717 examples respectively. We used their test set for testing, while all other examples were used for training.

References

1. Caselli, T., Vossen, P.: The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In: Proceedings of the Events and Stories in the News Workshop. pp. 77–86. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/W17-2711, https://aclanthology.org/W17-2711
2. Dunietz, J., Levin, L., Carbonell, J.: The BECauSE corpus 2.0: Annotating causality and overlapping relations. In: Proceedings of the 11th Linguistic Annotation Workshop. pp. 95–104. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-0812, https://aclanthology.org/W17-0812
3. Gao, L., Choubey, P.K., Huang, R.: 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). pp. 1808–1817.

[11] Belief: One argument gives the evidence justifying the claim in the other argument; SpeechAct: One argument provides the reason for the speaker uttering the speech act in the other argument [12]

[12] The first argument is Cause span for: Contingency.Cause.Result, Contingency.Purpose.Arg1-as-goal, Contingency.Condition.Arg1-as-cond, Contingency.Negative-condition.Arg1-as-negCond, and Contingency.Negative-cause.NegResult

[13] The second argument is Cause span for: Contingency.Cause.Reason, Contingency.Purpose.Arg2-as-goal, Contingency.Condition.Arg2-as-cond, and Contingency.Negative-condition.Arg2-as-negCond

[14] Dev set: wsj_00, wsj_01, wsj_24 ; Test set: wsj_22, wsj_23

Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1179, https://aclanthology.org/N19-1179
4. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), https://aclanthology.org/S10-1006
5. Hidey, C., McKeown, K.: Identifying causal relations using parallel Wikipedia articles. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1424–1433. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1135, https://aclanthology.org/P16-1135
6. Mirza, P., Sprugnoli, R., Tonelli, S., Speranza, M.: Annotating causality in the TempEval-3 corpus. In: Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL). pp. 10–19. Association for Computational Linguistics, Gothenburg, Sweden (Apr 2014). https://doi.org/10.3115/v1/W14-0702, https://aclanthology.org/W14-0702
7. Mirza, P., Tonelli, S.: An analysis of causality between events and its relation to temporal information. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2097–2106. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (Aug 2014), https://aclanthology.org/C14-1198
8. Mirza, P., Tonelli, S.: CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 64–75. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://aclanthology.org/C16-1007
9. Ponti, E.M., Korhonen, A.: Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics. pp. 25–30. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-0903, https://aclanthology.org/W17-0903
10. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A.K., Webber, B.L.: 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 (2008), http://www.lrec-conf.org/proceedings/lrec2008/summaries/754.html
11. Tan, F.A., Hürriyetoğlu, A., Caselli, T., Oostdijk, N., Nomoto, T., Hettiarachchi, H., Ameer, I., Uca, O., Liza, F.F., Hu, T.: The causal news corpus: Annotating causal relations in event sentences from news. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 2298–2310. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.246
12. Webber, B., Prasad, R., Lee, A., Joshi, A.: The penn discourse treebank 3.0 annotation manual. Philadelphia, University of Pennsylvania (2019)
13. Zuo, X., Cao, P., Chen, Y., Liu, K., Zhao, J., Peng, W., Chen, Y.: Improving event causality identification via self-supervised representation learning on external causal statement. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2162–2172. Association for Computational Lin-

guistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.190, https://aclanthology.org/2021.findings-acl.190