A New Dataset for Causality Identification in Argumentative Texts

Anonymous ACL submission

Abstract

Existing datasets for causality identification in argumentative texts have several limitations, such as the type of input text (e.g., only claims), causality type (e.g., only positive), and the linguistic patterns investigated (e.g., only verb connectives). To resolve these limitations, we build a new dataset with sophisticated inputs (all units from arguments), a balanced distribution of causality types, and a larger number of linguistic patterns denoting causality. To this end, we combine the two paradigms of distant supervision and uncertainty sampling to identify diverse, high-quality samples of causality relations, and annotate them cost-effectively.

1 Introduction

Causality identification is a vital task in natural language processing that can contribute to different downstream applications such as question answering, fact-checking, and commonsense reasoning. The task which concerns identifying texts with causality relations, the type of relations (positive or negative), and the concepts involved in the relations, is studied in diverse domains including biomedicine (Kyriakakis et al., 2019), education (Stasaski et al., 2021), and recently computational argumentation (AlKhatib et al., 2020).

In computational argumentation, causality identification impacts fundamental tasks such as topic-independent argument mining and large-scale argumentation graphs constructing (Reisert et al., 2018). Despite its importance, only a few annotated datasets for identifying causality have been built so far. Moreover, these datasets often focus only on one argument component (e.g., claim), encode bias towards the 'positive' type of causality, and/or consider a limited number of linguistic patterns that capture causality (e.g., verb connectives). As such, developing *robust* supervised learning approaches based on these datasets for causality identification becomes more laborious.

This paper aims to expand and enrich the available data for causality identification in argumentative texts written in English with a new dataset that deals with more sophisticated input text (i.e, the whole argument), covers more causality patterns, and maintains a balanced distribution of causality types. To this end, we develop an approach that comprises two main steps: *distant supervision* and *uncertainty sampling*. The former resulting in 10,329 candidate sentences for causality, and the second in 1,485 manually annotated argumentative sentences, 867 of which contain at least one causal relation. Of these, 515 sentences are further annotated as containing a positive cause-effect relation, and 536 as containing a negative one. Many sentences encode multiple relations, and involve diverse linguistic patterns (see Section 3).

We train transformer-based classifiers using our newly built dataset, reaching high effectiveness in identifying causal relations compared to several baselines. The developed resources in the paper (e.g., data and code) will be publicly available.[1]

2 Causality Dataset Construction

In this section, we describe our method for constructing the causality dataset. In particular, we first outline the distant supervision step, then, we discuss the uncertainty sampling.

2.1 Distant Supervision

Distant supervision is the process of mining suitable training examples from weakly labeled data sources using task-specific heuristics (Mintz et al., 2009). These examples can then be used for supervised learning of the task in hand. Here, we employ distant supervision to find argumentative sentences that probably encode causality relations, without being restricted to certain topics or linguistic patterns. Specifically, we first collect pairs of

[1]The full annotation set is in the supplementary material.

concepts that are involved in a causality relation, using the corpus of AlKhatib et al. (2020). Second, we acquire a set of argumentative texts and segment them into self-contained sentences. Lastly, we extract the argumentative sentences that contain at least one of the concept pairs.

Concept Pairs Collection In this step, we utilize the corpus of AlKhatib et al. (2020) to collect various concepts related to causality. The corpus covers 4,740 claims extracted from Debatepedia – an online debate portal. Each claim is manually annotated for the presences of a causality relation (called 'effect'), the type of the relation (positive or negative), and the concepts that are involved in the found relation.

We carefully review these concepts and perform two filtering steps: (1) we simplify the complex concepts (e.g., from *"the state can regulate the sale"* to *"sale regulation"*), and we (2) split some concepts into multiple ones (e.g., *"crime and safety problems"* is split into *"crime"* and *"safety problems"*). Overall, we end up with 1,930 unique concepts grouped into pairs, each of which consists of two concepts involved in the same relation (e.g., *"legalizing marijuana"* and *"safety"*).

Argumentative Data Acquisition and Simplification We rely on the Args.me corpus (Ajjour et al., 2020) as the source of the argumentative data. The corpus includes 387,606 arguments from various debates regarding controversial topics. The arguments are derived from four popular debate portals: Debatewise (14,353 arguments), IDebate.org (13,522 arguments), Debatepedia (21,197 arguments), and Debate.org (338,620 arguments).

To split the arguments into coherent and self-contained sentences, we use Graphene (Cetto et al., 2018), an open information extraction tool. This tool performs discourse simplification, in which an input sentence is syntactically simplified and split (if necessary) into sentences with resolved co-reference and high coherence. Altogether, the arguments are segmented into 10,720,451 sentences.

Concept Pairs and Argumentative Data Matching In this step, for each sentence in the acquired argumentative data, we check whether it includes any of the concept pairs. Using full string matches between the concepts and the sentences' tokens (after stemming with Porter Stemmer (Porter, 1980)), we obtain around 28,000 sentences that match at least one concept pair. We additionally filter them by removing duplicates, all hyperlinks and special characters contained in the sentences. Besides, on manual inspection, we observed that matching with generic concepts such as *"individuals"* or *"corporation"* lead to noisy sentences not actually containing any causality relations and were therefore excluded. As a result, we end up with 10,329 sentences. To evaluate the filtering process, we check a random sample of 100 sentences before and after filtering. We observe an increase in the number of sentences with causality relations (from 56 to 70).

2.2 Uncertainty Sampling

Uncertainty sampling is one of the strategies employed in active learning (Settles, 2012). Given an initial classification model and a pool of unlabeled samples, the goal is to select those samples for labeling for which the classifier's confidence is lowest, i.e., the predicted class distribution is closest to uniform, and thus maximize the information gain to the model. Following this idea, we train causality identification models on the labeled samples in the (AlKhatib et al., 2020) dataset, and use the argumentative sentences acquired from the distant supervision step as the unlabeled pool. Next, based on the confidence of these models, we sample a subset of the sentences and annotate them manually via crowdsourcing.

Candidate Sentence Selection Causality identification is often comprised of three classification sub-tasks; given an input text, (1) detect whether the text contains a causality relation, (2) identify the type of causality, and (3) determine the entities or events representing the cause and effect relation.

For the first two sub-tasks, we develop several classification models,[2] using the corpus of AlKhatib et al. (2020), reaching high effectiveness.[3] We apply two of the best performing transformer-based models (RoBERTa and XLNet) to the 10,329 argumentative sentences obtained from the distant supervision step. Using these models' confidence scores, we distribute the sentences into nine bins: the first bin represents the highest

[2] The models are based on XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019), BERT (Devlin et al., 2018), NB-SVM (Wang and Manning, 2012), and Fasttext (Joulin et al., 2016). The implementation is done using the HuggingFace library (Wolf et al., 2020) with default settings.

[3] RoBERTa and XLNet achieve 0.88 and 0.91 F_1 scores for the first and second tasks respectively, compared to 0.81 and 0.86 achieved by the rich features SVM approach of (AlKhatib et al., 2020)

	Causality	Positive	Negative	Multiple
Expert	0.34	0.66	0.70	0.28
Crowd	0.27	0.31	0.36	0.03

Table 1: Inter-annotator agreement (Krippendorff's alpha) for the expert and crowdsourcing annotations.

	Expert		Crowd	
Causality				
Overall	80	100%	1324	100%
Relation	48	60%	819	62%
No Relation	32	40%	505	38%
Relation Type				
Overall	48	100%	819	100%
Positive	29	60%	486	59%
Negative	29	60%	507	62%
Multiple Relations				
Overall	48	100%	819	100%
Single	34	71%	614	75%
Multiple	14	29%	205	25%

Table 2: Sentence statistics of the new causality dataset. Relation type percentages do not sum to 100 since sentences can have multiple relations.

confidence for 'no-causality' class, and the last bin represents confidence for the 'causality' class.

Our aim is to find sentences that encode new causality patterns while maximizing the number of sentences with the 'negative-causality' class. Thus, our uncertainty sampling filters out the sentences with high confidence for 'causality', 'no-causality', and 'positive-causality' classes. This results in 1,937 sentences for our manual annotation.

Sentence-level Manual Annotation We conduct an annotation task for causality identification via Amazon Mechanical Turk for the 1,937 sampled sentences,[4] which requires identifying all the causality relations in a sentence. In particular, for each identified causality relation, the workers are asked to specify the causality relation's type, the concepts involved in the relation, and the sentence's phrase(s) that indicate the presence and type of the relation. The workers also have the option to point out the sentences that are not comprehensible or/and include several grammatical errors. [5]

We first ask three experts in computational linguistics to annotate 100 sentences, and use their feedback to refine the guideline and improve the annotation interface for the crowdsourcing task. Each sentence is annotated by three different workers. For quality control, we hire native English speakers with a task approval rate of at least 98%. We closely monitor and review the annotations, rejecting workers that perform poorly. In total, 285 workers successfully participate in our task, resulting in 1,485 sentences with high-quality annotations.

3 Causality Dataset Analysis

3.1 Qualitative Analysis

Inter-annotator Agreement The inter-annotator agreement, measured using Krippendorff's alpha and presented in Table 1, provides insights into the level of agreement among both experts and crowds. While the crowd's agreement is relatively lower compared to experts, they still achieve a reasonable level of agreement for causality and types (ranging from 0.27 to 0.36). However, the crowd tends to prioritize annotating only one relation per sentence, potentially overlooking instances with multiple relations. These findings highlight the subjective nature of the task and the intricate linguistic patterns within the sentences. It is worth noting that the majority of cases fall into the scenario where two out of the three annotators agree, which significantly helps in obtaining a reliable gold standard.

Dataset Statistics The annotations are aggregated based on majority vote, with one exception: we consider a sentence to have multiple relations as long as at least one annotator found multiple relations there. Table 2 shows statistics for the new causality dataset: there is a high percentage of causal relations, especially of the negative type; a quarter of the sentences contain more than one relation. This demonstrates the cost-effectiveness of our construction method; we obtain a rich set of causal sentences by annotating only 1,937 examples. The annotation study costs around 400 EUR.

Dataset Inspection We manually examine the dataset, exploring the causality linguistic patterns and the structure of the sentences with multiple relations. As for the linguistic patterns, we look at the list of phrases (provided by the annotators) that indicate a causality relation, finding different causal connectives such as verbs (*"prevent, "promote"*), verb phrases (*"leads to"*), conjunctions (*"because"*), prepositional phrases (*"because of, due to"*), and clauses (*"the source of, is an addition to, can be tied to, becomes a burden for"*).

[4] We pay a fair hourly wage for the annotators.

[5] The task instructions are carefully explained using written guidelines and demonstration videos, covering various causality relations with different linguistic patterns.

$\mathbf{X} \xmapsto{\text{positive}} \mathbf{A, B}$ Social media[X] can fuel anxiety[A] and depression.[B]
$\mathbf{X} \xmapsto{\text{negative}} \mathbf{A, C, D}$; $\mathbf{X} \xmapsto{\text{positive}} \mathbf{B}$ GM foods[X] are safe for human consumption, reduce pesticide,[A] increase yield,[B] decrease cost,[C] and combat global warming.[D]
$\mathbf{X} \xmapsto{\text{negative}} \mathbf{A, B, C, D}$; $\mathbf{Y} \xmapsto{\text{positive}} \mathbf{A, B, C, D}$ Marijuana[X] can relieve certain types of pain,[A] nausea,[B] vomiting[C] and other symptoms[D] caused by such illnesses as cancer.[Y]
$\mathbf{X} \xmapsto{\text{negative}} \mathbf{Y}$; $\mathbf{X} \xmapsto{\text{positive}} \mathbf{<Z} \xmapsto{\text{negative}} \mathbf{Y>}$ Genetic screening[X] for the embryos can reduce the chance of giving birth to more than one child;[Y] because clinics[Z] now want to prevent this by planting one embryo at a time and they have to do this through genetic screening.

Table 3: Examples of the found patterns for causality in the set of the sentences with multiple relations.

Classifier	D_1	D_2	D_1+D_2
C_1	0.88	0.63	0.82
C_2	0.74	0.71	0.74
C_3	**0.89**	**0.75**	**0.85**
Majority Class Baseline	0.64	0.53	0.62
(AlKhatib et al., 2020)	0.81	-	-

Table 4: F_1 scores for causality identification. D_1 is the test set of AlKhatib et al. (2020), D_2 is our test set.

Besides, we find different patterns for causality in the sentences that contain multiple relations. Examples of these patterns are shown in Table 3.

3.2 Quantitative Analysis

To evaluate the impact of our constructed dataset, we employ it to develop a new classifier for causality identification. We compare the effectiveness of this classifier to another one that is developed using the corpus of AlKhatib et al. (2020).

In particular, we tackle the tasks of identifying whether a sentence has causality relation(s),[6] implementing three classifiers based on the XLNet model: (C_1) this classifier is trained by the training set of AlKhatib et al. (2020), (C_2) this classifier is trained by 80% of our new constructed dataset, and (C_3) this classifier is trained by the combination of the training sets of the new and old classifiers.

We apply the three classifiers to the test set of AlKhatib et al. (2020) (D_1), the test set of our new dataset (20%) (D_2), and both test sets combined (Table 4). In general, the classifier trained on (D_2) outperforms the baseline, and using the classifier that is trained with the combined training set (C_3) always leads to the best effectiveness, which speaks for the positive impact of our new dataset.

4 Related Work

In general, causality datasets are expensive to build, scarce, small, biased towards one class, focused on only a single aspect of causality (e.g., whether a sentence has a causal relation or not), and include limited linguistic patterns (due to their sampling method, e.g., via a seed list of causal verbs). Recently, Xu et al. (2020) reviewed six publicly-available datasets. The largest, Altlex (Hidey and McKeown, 2016), comprises nearly 45,000 sentences, but they are annotated only for the presence of causal relations, and only 10% are causal; the other five datasets are an order of magnitude smaller, and exhibit similar bias.

In addition, the EventStoryLine Corpus (Caselli and Vossen, 2017), which is frequently used in related work, comprises several thousand causal links but no annotated negative samples. AlKhatib et al. (2020) present a corpus of 4,740 claims from argumentative texts, 36% of which are annotated as containing a causal relation.

Perhaps most closely related to our own work, Zuo et al. (2020) propose a distant supervision-based data augmentation framework to address the data scarcity problem in causality. Whereas their approach involves a fully automated causal event pair extraction for distant supervision, we propose a human-in-the-loop framework based on uncertainty sampling, aiming to both improve the quality, and drive down the cost of hand-labeled corpora.

5 Conclusion

In this paper we present a new dataset for causality identification in argumentative texts that considers all argument units (claims, premises) as inputs. The sentences in our resulting dataset comprise a balanced distribution of causality types and diverse linguistic patterns denoting causality. Initial experiments on causality identification using transformer-based classifiers demonstrate the effectiveness of our smaller yet high-quality dataset in comparison to a larger existing corpus with some limitations. In future, we plan to employ the new dataset for mining causality relations from the Web.

[6] We focus on causality identification because AlKhatib et al. (2020) do not consider multiple relations, making their dataset partially incompatible for causality type identification.

References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2020. args.me corpus.

Khalid AlKhatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7367–7374. AAAI Press.

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. Association for Computational Linguistics.

Matthias Cetto, Christina Niklaus, André Freitas, and Siegfried Handschuh. 2018. Graphene: Semantically-linked propositions in open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2300–2311. Association for Computational Linguistics.

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*.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Manolis Kyriakakis, Ion Androutsopoulos, Artur Saudabayev, and Joan Ginés i Ametllé. 2019. Transfer learning for causal sentence detection. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 292–297. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers.

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A. Hearst. 2021. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170, Online. Association for Computational Linguistics.

Sida I. Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 90–94. The Association for Computer Linguistics.

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.

Jinghang Xu, Wanli Zuo, Shining Liang, and Xianglin Zuo. 2020. A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th*

International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, pages 1519–1531. International Committee on Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. Knowdis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1544–1550. International Committee on Computational Linguistics.