Zero-shot Event Causality Identification with Question Answering

Daria Liakhovets
AIT Austrian Institute of Technology
daria.liakhovets@ait.ac.at

Sven Schlarb
AIT Austrian Institute of Technology
sven.schlarb@ait.ac.at

Abstract

Extraction of event causality and especially implicit causality from text data is a challenging task. Causality is often treated as a specific relation type and can be considered as a part of relation extraction or relation classification task. Many causality identification-related tasks are designed to select the most plausible alternative of a set of possible causes and consider multiple-choice classification settings.

Since there are powerful Question Answering (QA) systems pretrained on large text corpora, we investigated a zero-shot QA-based approach for event causality extraction using a Wikipedia-based dataset containing event descriptions (articles) and annotated causes. We aimed to evaluate to what extent reading comprehension ability of the QA-pipeline can be used for event-related causality extraction from plain text without any additional training. Some evaluation challenges and limitations of the data were discussed. We compared the performance of a two-step pipeline consisting of passage retrieval and extractive QA with QA-only pipeline on event-associated articles and mixed ones. Our systems achieved average cosine semantic similarity scores of 44 – 45% in different settings.

Keywords: event causality identification, question answering, semantic similarity search.

1 Introduction

The aim of the work was to exploit the reading comprehension of pre-trained Question Answering (QA) models to address zero-shot event causality extraction from text. Since implicit causality can be expressed in various, potentially infinite number of ways, and causality expressions can be distributed throughout sentences, identification of event causality remains a challenging task.

Many related data resources are designed for binary statement classification, multiple-choice QA, or relation classification. For our experiments we used a semantic similarity search-based dataset obtained from annotated Wikipedia articles. The dataset was designed for event-related causality extraction from plain text. However, the data had some limitations discussed in the related section.

We compared a two-step extraction pipeline consisting of relevant text passage retrieval based on semantic similarity search and cause candidate retrieval based on QA. The experiments were performed in two different settings: related documents and mixed documents subsets.

The paper is structured as follows: Section 2 overviews related work on causality identification, including some question-driven approaches. Section 3 describes our data, experiments and evaluation metrics, and Section 4 presents the results.

2 Related work

2.1 Causality identification: resources and approaches

Resources, approaches, and problems in causal relation identification in NLP are discussed by (Han and Wang, 2021). The authors distinguish causal relation classification and causal relation extraction and the classification level (word-, sentence- or passage-level). Causality is often treated as a specific type of entity relations. Some datasets combine event causality and temporal relations, e.g., (Caselli and Vossen, 2017). There are some domain-specific resources, e.g., (Kyriakaki et al., 2019), (Mariko et al., 2020). Others, e.g., (Huang et al., 2019), (Ponti et al., 2020), are designed for commonsense multiple-choice causal QA. There are also knowledge bases containing causal relations or lexical markers.

Causality expressions can be explicit (e.g., “because”) or implicit, the latter are more common but more difficult to recognize. Open class lexical markers, AltLexes (Prasad et al., 2008), are somewhere in the middle due to their linguistic variety (Hidey and McKeown, 2016).

Since existing labelled event causality detection datasets are limited in size, data augmentation techniques used, such as synonym substitution (Staliūnaitė et al., 2021) or external causal knowledge (Dalal et al., 2021). (Zuo et al., 2020) suggested a data augmentation framework based on lexical and causal commonsense knowledge. (Ruan et al., 2019) used WHY-type question-answer pairs from QA datasets and Question-Statement Conversion for training set expansion.

(Han and Wang, 2021) summarize methods for causal relation identification. While unsupervised methods are mainly based on predefined rules and patterns, supervised methods use feature engineering, global optimization, or deep learning approaches on labelled data. Despite the achieved good performance in many causal relation identification tasks, extracting implicit causal knowledge from the free text is still an unsolved task.

(Doan et al., 2019) used dependency parsing on lemmatized POS-tagged tweets to extract cause-effect relations for several health-related effects (e.g., “headache”). (Kyriakaki et al., 2019) used transfer learning to detect causal sentences in commonsense datasets and in BioCausal data and experimented with the BIGRUATT layer. (Kadowaki et al., 2019) investigated ensemble approaches based on individual judgements of three annotators and exploiting background context knowledge for binary classification of statement pairs. (Mariko et al., 2020) fine-tuned BERT for binary sentence classification in financial news. (Liang et al., 2022) proposed a novel model that exploits the advantages of both feature engineering and neural model-based approaches. (Zhao et al., 2021) proposed a document-level context-based graph inference mechanism to identify event causality.

2.2 Question-driven approaches

Event causality identification can be considered as a part of automated story generation. (Castricato et al., 2021) proposed a novel approach that reconstructs the story backwards by iteratively generating “why“-questions to find the preceding event from the given one. (Zhou et al., 2021) used QA to identify nested causality in traffic accident data.

Zero-shot methods aim to overcome the limitations of predefined relation set-based approaches towards extracting new unseen types of relations or facts. (Levy et al., 2017) used QA to perform zero-shot relation extraction by associating natural-language questions with each relation type and demonstrated the generalization ability of the approach on unseen relation types. (Goodwin et al., 2020) applied multi-task fine-tuning for zero-shot conditional summarization that selects the most salient points based on a question or a topic of interest. (Chakravarti et al., 2020) addressed a zero-shot industrial QA task introducing the model GAAMA with improved attention mechanisms. (Zhou et al., 2021) proposed a novel method for automatic transfer of explanatory knowledge in zero-shot science QA.

3 Experimental setup

3.1 Data

To address event-related causality identification from free text, we obtained a dataset from the Wikipedia *List of protests in the 21st century*[1]. The dataset language was English. We extracted human-annotated “caused by” attributes from “infobox” sections (Figure 1).

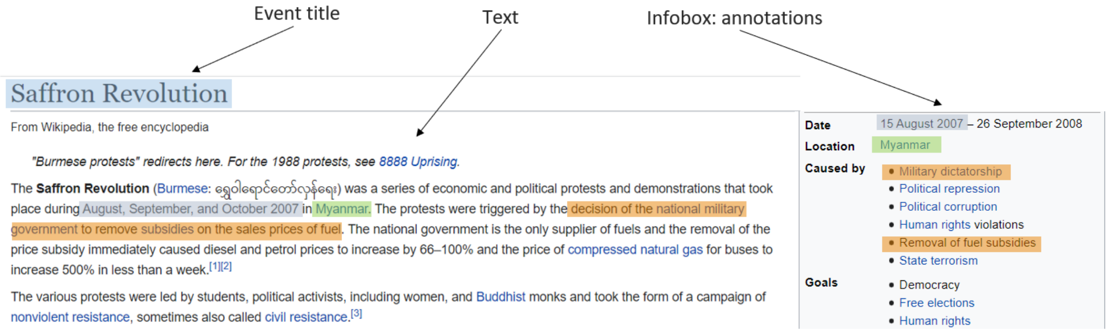
Since extractive methods require annotations to appear in text, we looked for annotated causes in text. Some annotations were matched exactly in the related article, others had to be searched for by their paraphrased appearances, e.g., *“authoritarianism”* could be found as *“authoritarian rule”*.

We created two dataset versions: using fuzzy string-matching functions from `thefuzz` [2] package and using semantic similarity search with `Sentence Transformers` [3] introduced by (Reimers and Gurevych, 2019). While the first

[1] https://en.wikipedia.org/wiki/List_of_protests_in_the_21st_century

[2] https://github.com/seatgeek/thefuzz

[3] https://www.sbert.net/



Saffron Revolution

From Wikipedia, the free encyclopedia

"Burmese protests" redirects here. For the 1988 protests, see 8888 Uprising.

The **Saffron Revolution** (Burmese: ရွှေဝါရောင်တော်လှန်ရေး) was a series of economic and political protests and demonstrations that took place during August, September, and October 2007 in Myanmar. The protests were triggered by the decision of the national military government to remove subsidies on the sales prices of fuel. The national government is the only supplier of fuels and the removal of the price subsidy immediately caused diesel and petrol prices to increase by 66–100% and the price of compressed natural gas for buses to increase 500% in less than a week.[1][2]

The various protests were led by students, political activists, including women, and Buddhist monks and took the form of a campaign of nonviolent resistance, sometimes also called civil resistance.[3]

Date	15 August 2007 – 26 September 2008
Location	Myanmar
Caused by	• Military dictatorship • Political repression • Political corruption • Human rights violations • Removal of fuel subsidies • State terrorism
Goals	• Democracy • Free elections • Human rights

Figure 1: Wikipedia-article with an infobox-section.

approach is based purely on token similarity, the second one uses embeddings produced by the `all-mpnet-base-v2`[4] model to compute the cosine similarity of two sequences. Thus, it can capture the semantic content, even if it is expressed with different words. As many annotated causes appear as paraphrased expressions, we used the second version of the dataset for our experiments. The minimum threshold of cosine similarity score was set to 0.70 to obtain a subset with better appearances of original annotations. The final subset contained 905 annotated causes linked to 297 unique articles; 245 causes were matched exactly (score 1.00), and 660 causes with similar phrases.

The data has the following limitations:

- **Objectivity:** authors of the Wikipedia articles may be biased. One may argue whether such annotations should be used as ground truth labels.
- **Completeness:** causal reasons may appear in the text without being annotated and therefore cannot be evaluated reliably.
- **Unlinked and inconsistently structured annotations:** firstly, annotations are not linked to their appearance in the text. For reliable evaluation, approximatively matched causes should be confirmed manually. Secondly, authors use different separators and list styles. Splitting the annotations into single cause items may break sentences into parts unevaluable for causality.

3.2 Question-driven cause candidate extraction

We used a question-driven two-step extraction approach to identify the cause candidates for an event of interest. To extract causality of a specific event, we constructed a question using the event title – in our case the Wikipedia article title – to complete the following simple question template:

What caused <EVENT_TITLE>?

We split articles into smaller passages, with a maximum of 200 `WordPiece` (Schuster and Nakajima, 2012) tokens, retaining the text structure, i.e., sentences and paragraphs. We exploited embeddings from `multi-qa-mpnet-base-dot-v1`[5], a model designed for semantic search to compute the dot similarity score of the question and passages and extract relevant ones (Figure 2).

In the next step we used `xlm-roberta-large-squad2`[6], a model designed for extractive QA, to retrieve answers from three most relevant passages (Figure 3). Since one article usually had multiple annotations, we retrieved several answer candidates from each passage and then selected two best ranked answers more than the number of annotations. Answer candidates were selected based on their probability of being an answer for the asked questions, which was calculated by the QA model.

Once several cause candidates for the article had been extracted, we had to match them with the annotations. We computed pairwise cosine similarity scores based on `all-mpnet-base-v2`

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[5] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

[6] https://huggingface.co/deepset/xlm-roberta-large-squad2

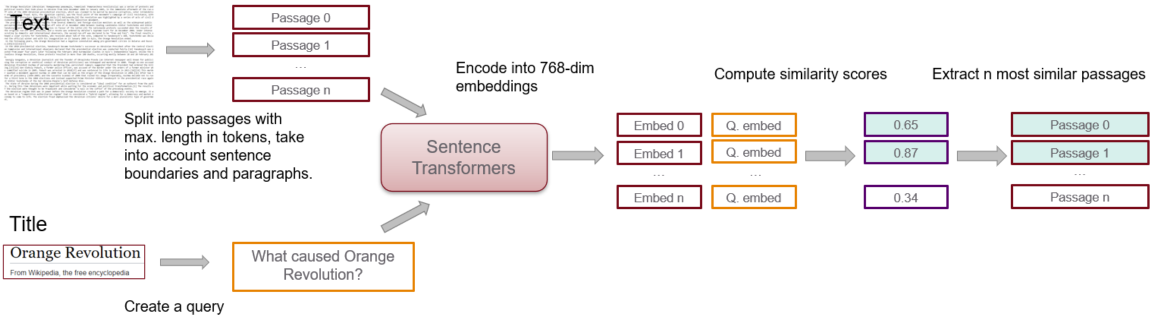

Text
Passage 0
Passage 1
...
Passage n
Split into passages with max. length in tokens, take into account sentence boundaries and paragraphs.
Title
Orange Revolution
From Wikipedia, the free encyclopedia
What caused Orange Revolution?
Create a query
Encode into 768-dim embeddings
Sentence Transformers
Compute similarity scores
Embed 0
Embed 1
...
Embed n
Q. embed
Q. embed
...
Q. embed
0.65
0.87
0.34
Extract n most similar passages
Passage 0
Passage 1
...
Passage n

Figure 3: Step 1: Passage retrieval using semantic similarity.

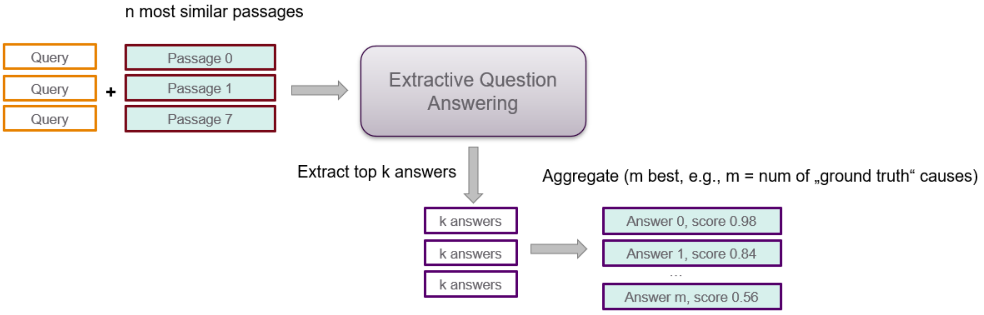


Figure 2: Step 2: Cause candidate extraction using QA.

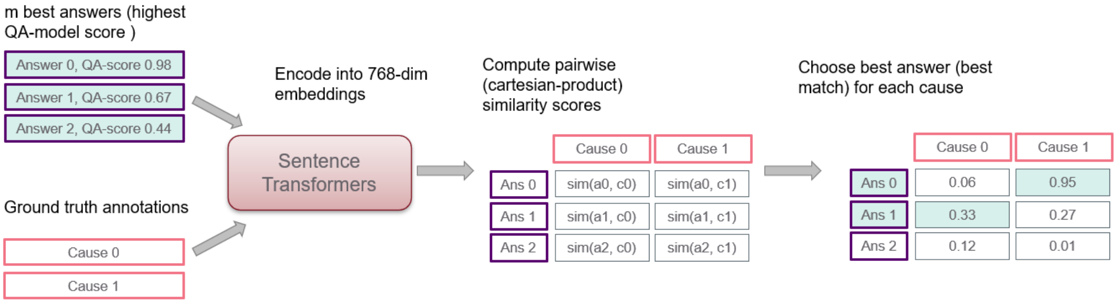


Figure 4: Matching annotations with extracted answers.

embeddings. For each annotation, the best match was selected and then used for the evaluation (Figure 4).

We compared the two-step extraction with the QA-only approach which has no passage retrieval step and just retrieves answer candidates from each text passage.

We experimented with two settings: extracting cause candidates only from a related article and from mixed documents, which is more realistic. In the second case, for each article we created a subset of 10 documents: the article itself and 9 random articles.

3.3 Evaluation metrics

To evaluate the retrieved answer candidates, we used the semantic similarity score (cosine similarity) computed based on `all-mpnet-base-v2` embeddings during best answer matching, F1-score, and exact match (EM). We removed punctuation and stop-words and compared two lowercased sets of tokens to obtain the F1-score. For EM, we compared two lowercased phrases without punctuation.

The F1-score is based on the lexical overlap of two token sequences, and EM just indicates whether the sequences are identical or not. Since more than 70% of entries in our data cannot be found exactly in the related articles, the semantic similarity score is more useful for evaluation. One could also use cross-encoder model-based scoring, as proposed by (Risch et al., 2021). For measuring lexical overlap, ROUGE metric (Lin, 2004) can be useful.

4 Results and discussion

Metric	Rel.: 2-step	Rel.: QA-only
Cos. similarity, avg.	0.4451	0.4588
F1-score, avg.	0.1516	0.1666
Exact match, n	7	9
No answer, n	0	0

Table 3: Evaluation results on related articles.

Metric	Mixed: 2-step	Mixed: QA-only
Cos. similarity, avg.	0.4397	0.4386
F1-score, avg.	0.1489	0.1513
Exact match, n	7	8
No answer, n	3	0

Table 2: Evaluation results on mixed articles.

The evaluation results are summarized in Table 1 (related articles) and Table 2 (mixed articles).

The number of exact matches is very low (< 1%) in all cases. There is no significant difference between the two approaches, judging by the metrics. However, QA-only is more time-intensive because it processes all text chunks.

The QA-only approach provided two and one more exact matches than passage retrieval + QA in related document- and mixed document-settings, respectively, as well as slightly higher F1-scores. In the mixed setting QA-only was able to find candidates for all annotations while the two-step approach missed candidates for three causes, i.e., some salient text passages were ignored during passage retrieval. This issue can be addressed by increasing the number of passages and/ or improving the quality of passage ranking techniques. However, we still think that the passage extraction step can have advantages when dealing with large text collections. Further experiments are needed to prove this.

The results could be improved by additional domain-specific model training and increasing the number of retrieved passages and answer candidates. Generative summarization could be a better choice than using only extractive methods.

Table 3 contains some examples of extracted cause candidates. The top half refers to the dataset: the "True cause" column contains original annotations, "Best match" presents the most similar phrase found in the article, and "Matching score" shows their similarity score. In the bottom half, "Best answer" contains the best candidate for the "True cause" and the appropriate "Answer matching score". "QA score" presents scores computed by the QA model.

#	Example 1	Example 2	Example 3
True cause	Mexican Drug War	2017 wealth tax repeal	religious nationalism
Best match	Mexican Drug War	to reinstate a wealth tax	nationalism
Matching score	1.00	0.76	0.85
Best answer	Mexican Drug War,	Their principal concern was tax justice.	mobs attacking Muslims.
QA score	0.02	0.22	0.79
Answer matching score	0.97	0.44	0.41

Table 1: Examples of results. Top half: True cause, Best match, Matching score refer to the dataset. Bottom half: Best answer, QA-score, Answer matching score refer to extracted causes.

The first example demonstrates a large gap between the low probability of being an answer to the asked question and the high score of matching with the ground truth annotation. In a real-world application, relying on the QA score, this answer would be low-ranked. The second example can be considered satisfactory by human judgement. Although the best answer conveys the main idea of the true annotation, its answer matching score is relatively low, as well as its QA score. The third example illustrates a cause candidate scored highly by the QA model but having a relatively low answer matching score. These examples demonstrate the need to define a sufficient level of similarity, because even similarity scores under 0.5 may still indicate adequate matches.

5 Conclusions and future work

In this work, we conducted experiments to evaluate the zero-shot event causality identification with semantic search-based passage retrieval and QA on a dataset obtained from Wikipedia. We compared the two-step and the QA-only approaches on related and mixed documents and demonstrated their similar performance in the experimental settings. While the two-step approach could not find any candidates for a few ground truth annotations in the mixed document setting, QA-only was able to find candidates in all cases. QA-only also performed slightly better on related documents, however, it required more computational time. Further experiments are necessary to identify whether the passage retrieval step bring other advantages when processing large document collections. Our systems achieved average cosine semantic similarity scores of 44 – 45% in different settings.

We think that the reading comprehension of QA models can be used to address the challenge of event causality extraction. In the future work, both passage and answer retrieval can be improved by using models with domain-specific knowledge, as well as increasing the number of retrieved passages and candidate. Using other or multiple question templates could help to retrieve more various cause candidates.

References

Caselli, T. & Vossen, P., 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. *Proceedings of the Events and Stories in the News Workshop*, p. 77–86.

Castricato, L., Frazier, S., Balloch, J. & Riedl, M., 2021. Tell Me A Story Like I'm Five: Story Generation via Question Answering. *Proceedings of the 3rd Workshop on Narrative Understanding*.

Chakravarti, R. et al., 2020. Towards building a Robust Industry-scale Question Answering System. *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*.

Dalal, D., Arcan, M. & Buitelaar, P., 2021. Enhancing Multiple-Choice Question Answering with Causal Knowledge. *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.

Doan, S. et al., 2019. Extracting health-related causality from twitter messages using natural language processing. *BMC Med Inform Decis Mak 19*.

Goodwin, T. R., Savery, M. E. & Demner-Fushman, D., 2020. Towards Zero-Shot Conditional Summarization with Adaptive Multi-Task Fine-Tuning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Han, M. & Wang, Y., 2021. A Survey on the Identification of Causal Relation in Texts. *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pp. 1-7.

Hidey, C. & McKeown, K., 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1424–1433.

Huang, L., Bras, R. L., Bhagavatula, C. & Choi, Y., 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. *EMNLP'2019*.

Kadowaki, K. et al., 2019. Event Causality Recognition Exploiting Multiple Annotators' Judgments and Background Knowledge. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Kyriakaki, M., I. A., Ametllé, J. G. i. & Saudabayev, A., 2019. Transfer Learning for Causal Sentence Detection. *BioNLP 2019 workshop*.

Levy, O., Seo, M., Choi, E. & Zettlemoyer, L., 2017. Zero-Shot Relation Extraction via Reading Comprehension. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*.

Liang, S. et al., 2022. A multi-level neural network for implicit causality detection in web texts. *Neurocomputing, Volume 481*, pp. 121-132.

Lin, C.-Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, p. 74–81.

Mariko, D. et al., 2020. The Financial Document Causality Detection Shared Task (FinCausal 2020). *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*.

Ponti, E. M. et al., 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Prasad, R. et al., 2008. The Penn Discourse TreeBank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Reimers, N. & Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* January, pp. 3973-3983.

Risch, J., Möller, T., Gutsch, J. & Pietsch, M., 2021. Semantic Answer Similarity for Evaluating Question Answering Models. *Proceedings of the 3rd Workshop on Machine Reading for Question Answering,* November, p. 149–157.

Ruan, H. et al., 2019. Using WHY-type Question-Answer Pairs to Improve Implicit Causal Relation Recognition. *2019 International Conference on Asian Language Processing (IALP)*.

Schuster, M., & Nakajima, K., 2012. Japanese and Korean Voice Search. *International Conference on Acoustics, Speech and Signal Processing,* p. 5149-5152.

Staliūnaitė, I., Gorinski, P. J. & Iacobacci, I., 2021. Improving Commonsense Causal Reasoning by Adversarial Training and Data Augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhao, K. et al., 2021. Document-level event causality identification via graph inference mechanism. *Information Sciences 561(3)*.

Zhou, G. et al., 2021. Nested Causality Extraction on Traffic Accident Texts as Question Answering. *NLPCC 2021: Natural Language Processing and Chinese Computing*.

Zhou, Z., Valentino, M., Landers, D. & Freitas, A., 2021. Encoding Explanatory Knowledge for Zero-shot Science Question Answering. *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*.

Zuo, X., Chen, Y., Liu, K. & Zhao, J., 2020. KnowDis: Knowledge Enhanced Data Augmentation for Event Causality Detection via Distant Supervision. *Proceedings of the 28th International Conference on Computational Linguistics*.