SemEval 2020 Task Proposal: Modelling Causal Reasoning in Language: Detecting Counterfactuals

Xiaodan Zhu[1], Xiaoyu Yang[1], Huasha Zhao[2], Qiong Zhang[2],
[1]Queen's University, Kingston, Canada
[2]Alibaba Group, San Mateo, USA
{xiaodan.zhu, xiaoyu.yang}@queensu.ca
{qz.zhang, huasha.zhao}@alibaba-inc.com

1 Introduction

Causal reasoning, a core constituent of intelligence, is used daily in human language. The most recent years have seen significantly increasing interest in modelling inference and reasoning in natural language (Zellers et al., 2018; Bowman et al., 2015; MacCartney and Manning, 2009). In this shared task, we focus on modelling causal reasoning.

Our task aims to provide a benchmark to evaluate the state-of-the-art models for causal reasoning in natural language. While causal reasoning is a broad topic, counterfactual is regarded as the top rung of the ladder of causality (Pearl and Mackenzie, 2018) and is one of the most active areas in recent years and has received intensive studies in different disciplines (Pearl and Mackenzie, 2018; Son et al., 2017; Kray et al., 2010; Buffone et al., 2016; Vandenbroucke et al., 2016) .

To model counterfactual semantics and reasoning in natural language, our shared task aims to provide a benchmark for two basic problems. Participants of subtask 1 (Section 2) are asked to determine whether a given statement is counterfactual or not. Counterfactual statements describe events that did not actually happen or cannot happen, as well as the possible consequence if the events have had happened (See Task Description of Section 2 for details). This problem is the most basic problem for all down-stream study on counterfactual-related inference in natural language. Subtask 2 (Section 3) further locates the antecedent and its consequent in a given counterfactual statement.

We believe the task will attract more and more attention and participation—if "new generation of robots should explain to us why things happened, and way they responded the way the did" (Pearl and Mackenzie, 2018), they are also expected to understand causal statements others say or write. Note that even our task focuses on modelling counterfactuals in text, it is related to multidisciplinary research, as counterfactual is studied in different fields. For example, thinking counterfactually could have an impact on human cognition, affections, and behaviors (Kray et al., 2010; Zeelenberg and Pieters, 2007). Research conducted in epidemiology investigates questions like *"which factors cause a certain disease?"*, and it is mainly about causal effects as well as counterfactual reasoning (Höfler, 2005; Vandenbroucke et al., 2016). By asking counterfactual questions, humans try to understand the significance of specific events in the big picture of life, and counterfactual is also regarded to be very important in evolution (Pearl and Mackenzie, 2018; Kray et al., 2010; Buffone et al., 2016). The landmark paper of Goodman (Goodman, 1947) gives a detailed analysis on counterfactual conditionals in philosophy and logistics.

For evaluation, we have collected 25,000 statements from news articles on three domains: finance, healthcare, and politics. We will describe the details of the tasks, evaluation, and baselines.

2 Subtask 1: Detecting Counterfactual Statements

Task Description Our first subtask asks participants to detect whether a given natural language statement is counterfactual or not. More specifically, counterfactuals describe events counter to facts and hence naturally involve common sense, knowledge, and reasoning. Tackling this problem is the basis for all down-stream counterfactual-related causal inference analysis in natural language. For example, the following statements are counterfactuals that need to be detected: one from healthcare and one from the finance domain.

- *Her post-traumatic stress could have been avoided if a combination of paroxetine and exposure therapy has been prescribed two months earlier.*
- *Finance Minister Jose Antonio Meade noted that if a jump in tomato prices had been factored out, inflation would have begun to drop.*

While the above examples are chosen for clarity for demonstration, real statements are much harder for computers to judge. (See Section 2 and (Son et al., 2017) for existing models and results).

Data Collection and Annotation Following (Son et al., 2017), we take a two-step strategy to collect and annotate the data.

Crawling candidate statements The first step is crawling news articles on the Web to obtain candidate counterfactual statements using the patterns listed in Appendix A. We develop our multi-domain counterfactual dataset for three domains: finance, politics, and healthcare. The URL we used to crawl news articles for each of these three domains are listed in Appendix B.

Our dataset is different from (Son et al., 2017) in the following aspects: (1) Ours is 10 times larger, which is critical for training complex models such as deep-learning-based models. However, the data developed earlier (Son et al., 2017) has only 1766 training and 1137 test statements. (2) Not only different in size, the data developed by (Son et al., 2017) are tweets that do not involve clear applications (e.g., tweets like *if all coffee shops played hip hop, I think the world would have been a better place.*). Our data are in three specific domains involving domain knowledge– these domains often use counterfactuals to express domain-specific causal reasoning (e.g., in the example we discussed above: *exposure therapy* may help cure *post-traumatic stress*). In addition, we expect the multi-domain data will help us understand counterfactual usage in different domains. (3) Our data are from news articles written with formal English. We hope the algorithms and studies focus more on modelling counterfactual itself instead of being distracted by "noises" in informal languages.

Counterfactual annotation In the above step, we have already acquired 25,000 candidate counterfactual statements and we aim to collect about 30,000, with each domain has 1/3 of the data. The next step is annotating the candidate statements to be either *counterfactual* or *non-counterfactual* on Amazon Mechanical Turk (AMT).

Each sample will be annotated by five different workers. We obtain the counterfactual label for each statement using majority voting. For our pilot study on 500 candidate statements, counterfactuals count for about 20% of the candidates. For the 30,000 statements, we expect to obtain approximately 6,000 counterfactual statements, with the number much larger than that in (Son et al., 2017), which only has around 900 true counterfactual statements.

We then split the data into 60:20:20 as training, development, and test set, respectively. In subtask 1, algorithms will be developed to perform classification to detect counterfactual statements from non-counterfactuals.

Pilot Study and Baseline In addition to the pilot study discussed above in data collection and annotation, we have also performed a pilot study to set up baselines. Specifically, we will provide as the baselines the majority voting and a hybrid pipeline classifier that uses both hand-crafted rules and a SVM classifier (based on features extracted from the statements). As shown in Table 1, on the tweet dataset collected in (Son et al., 2017), the hybrid SVM model achieves a better performance than that of a rule-based model or an SVM model based only on features automatically extracted from the statements. We will provide these models as the baselines for our shared task.

It would be very interesting to understand how the recent advance on deep leaning models can help solve the counterfactual detection problem, particularly as the task provides a relatively large dataset for training complex models. We expect SemEval participants and other researchers to develop such models with our dataset.

Table 1: Pilot results on the tweet counterfactuals collected by (Son et al., 2017)

	Precision	Recall	F1
Hybrid Model	0.71	0.84	0.77
Rules only	0.59	0.91	0.71
SVM	0.24	0.91	0.38

Evaluation metrics The evaluation metric used in subtask 1 will be precision, recall, and F1.

3 Subtask 2: Detecting Antecedent and Consequent

Task Description Indicating causal insight is an inherent characteristic of counterfactual. To further detect the causal knowledge conveyed in counterfactual statements, subtask 2 aims to locate antecedent and consequent in counterfactuals.

According to (Goodman, 1947), a counterfactual statement can be converted to a contrapositive with a true antecedent and consequent. Consider the "post-traumatic stress" example discussed above; it can be transposed into *"because her post-traumatic stress was not avoided, (we know) a combination of paroxetine and exposure therapy was not prescribed"*. Such knowledge can be not only used for analyzing the specific statement, but also be accumulated across corpora to develop domain causal knowledge (e.g., a combination of paroxetine and exposure may help cure post-traumatic stress).

Data Collection and Annotation As the goal of subtask 2 is to identify antecedent and consequent in a given counterfactual statement, the annotation will only be performed on statement that is labelled by human annotator as counterfactual. The annotation of subtask 2 is to ask workers to mark these two counterfactual components. If one of them is missing (e.g., if a counterfactual statement misses consequent), the corresponding mark will be absent/empty. Same as in some previous SemEval tasks with multiple subtasks, our subtask 2 will run right after the subtask 1 phase finishes.

Baselines The first baseline is a sequence labelling model based on bag-of-words features. Same as in name entity recognition, this baseline model will annotate the antecedent and consequent using the B/I/O scheme, denoting if a word is at the Beginning, Inside, or Outside an antecedent and consequent. Specifically, we tag each token in the sentence with **B-Ant, I-Ant, B-Con, I-Con**, or **O**. The second baseline uses the pointer network proposed in (Vinyals et al., 2015), which is a widely used neural-network model that can mark the start and end token of a text span under concern, after being trained with the corresponding data (here antecedent or consequent spans).

Other Annotation Note that to provide a reference for future multidisciplinary research, we also annotate more detailed information as discussed in Appendix C.

Evaluation As a boundary detection task, we adopt two metrics used in NER and Question Answering to evaluate subtask 2. The **Exact Match** requires the prediction of antecedent or consequent boundaries to match the gold-standard boundary exactly. The **F1 score** measures the average overlap the prediction and the ground truth.

4 Data Availability and Copyright

We do not have copyright issue. Please refer to Appendix D for the detailed discussion.

5 Task organizers

Xiaodan Zhu has previously co-organized **SemEval-2016 Task 6**. He has also competed for **SemEval-2013** and **SemEval-2014** as a main contributor of several top-ranking systems. Xiaodan Zhu's recent research interests include natural language inference, commonsense reasoning (e.g., Wingograd Schema Challenge), sentiment analysis, and financial text analytics. He is currently an assistant professor of Queen's University, Canada. He serve as area chairs for ACL (2019, 18), NAACL (2019), and COLING (2018); publication co-chair for COLING-2018; workshop co-chair for COLING-2020; co-chair for SemEval-2019 and 2020.

Xiaoyu Yang is an MSc student of Queen's University, under the supervision of Prof. Xiaodan Zhu. Her research interests are natural language inference and common sense reasoning.

Huasha Zhao is a Staff Research Scientist and Technology Lead at Alibaba Group. He received his MSc and PhD from EECS at University of California, Berkeley, and B.Eng. in Electrical Engineering from Tsinghua University. His research interests include information extraction, business intelligence, natural language processing, and machine learning. He chaired and organized the E-Commerce workshop at SIGIR in both 2017 and 2018, and served as PC members for many top machine learning conferences/journals.

Qiong Zhang is currently a Senior Staff Engineer at Alibaba's DAMO Academy. He holds a Ph.D. in computer science from Zhejiang University and was a researcher at the University of Wisconsin-Madison. His current research focuses on text mining and applications, including information extraction, knowledge discovery, and natural language interface.

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Anneke Buffone, Shira Gabriel, and Michael Poulin. 2016. There but for the grace of god: Counterfactuals influence religious belief and images of the divine. *Social Psychological and Personality Science*, 7(3):256–263.

Nelson Goodman. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy*, 44(5):113–128.

M Höfler. 2005. Causal inference based on counterfactuals. *BMC medical research methodology*, 5(1):28.

Laura J Kray, Linda G George, Katie A Liljenquist, Adam D Galinsky, Philip E Tetlock, and Neal J Roese. 2010. From what might have been to what must have been: Counterfactual thinking creates meaning. *Journal of personality and social psychology*, 98(1):106.

Bill MacCartney and Christopher D Manning. 2009. *Natural language inference*. Citeseer.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.

Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 654–658.

Jan P Vandenbroucke, Alex Broadbent, and Neil Pearce. 2016. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International journal of epidemiology*, 45(6):1776–1786.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.

Marcel Zeelenberg and Rik Pieters. 2007. A theory of regret regulation 1.0. *Journal of Consumer psychology*, 17(1):3–18.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

A Patterns in Data Collection

The patterns we use are listed below: *if...then, if...(had / had not / hadn't), would / could / should / might / ought / may (have / not have / haven't), wouldn't / couldn't / shouldn't have, if...(were / were not / weren't), Had..., Were..., Should..., what if, but for, if only, so long as, whether, rather, assume, assuming, suppose, supposing, provided, providing, imagine, imagining, conjure, conjuring, visualize, visualizing, conceptualize, conceptualizing, envisioning, envision, wish and unless*.

B The Websites of News Reports

We crawl news articles from different websites of three domains:
(1) ***finance***: cnbc.com, businessinsider.com, investing.com
(2) ***healthcare***: webmd.com
(3) ***politics***: politics news from thisisinsider.com and uk.reuters.com

C Other Annotation

For true counterfactual statements, we will annotate with more detailed information from two respects. One is the *direction* of counterfactuals, including "upward", "downward", and "not clear". The other is judgement of the rationality of the causal relations between antecedent and consequent part mentioned in the counterfactual statement based on common sense.

D Data Availability and Copyright

According to Section 107 of the Copyright Law [1], and 28A and 30 of the Copyright Acts [2], there is one exception to copyright infringement which is fair use (or fair dealing). Fair use is appropriate for public benefit purposes, like research. Our use is not of commercial nature and we only distribute several sentences extracted from news reports online for the purpose of research. Besides, we only use texts that are publicly available, and the source will be stated according to law.

[1] https://www.copyright.gov/title17/92chap1.html#107

[2] https://www.gov.uk/government/publications/copyright-acts-and-related-laws