Detecting Causal Language Use in Science Findings

Bei Yu[1], Yingya Li[1], and Jun Wang[2]

[1]School of Information Studies, Syracuse University
[2]Independent Researcher
{byu, yli48}@syr.edu, junwang4@gmail.com

Abstract

Causal interpretation of correlational findings from observational studies has been a major type of misinformation in science communication. Prior studies on identifying inappropriate use of causal language relied on manual content analysis, which is not scalable for examining a large volume of science publications. In this study, we first annotated a corpus of over 3,000 PubMed research conclusion sentences, then developed a BERT-based prediction model that classifies conclusion sentences into "no relationship", "correlational", "conditional causal", and "direct causal" categories, achieving an accuracy of 0.90 and a macro-F1 of 0.88. We then applied the prediction model to measure the causal language use in the research conclusions of about 38,000 observational studies in PubMed. The prediction result shows that 21.7% studies used direct causal language exclusively in their conclusions, and 32.4% used some direct causal language. We also found that the ratio of causal language use differs among authors from different countries, challenging the notion of a shared consensus on causal language use in the global science community. Our prediction model could also be used to help identify the inappropriate use of causal language in science publications.

1 Introduction

Establishing causality is one of the most important goals and concerns of science. Therefore, the language that describes causal relationships plays a crucial role in communicating science findings among scientists and with the general public (Kleinberg and Hripcsak, 2011). However, scientists and journalists have been found to inappropriately use causal language in research publications and news articles. Specifically, causal interpretation of correlational findings from observational studies has been a major type of misinformation in science communication (Cofield et al., 2010; Sumner et al., 2014; Chiu et al., 2017; Boutron and Ravaud, 2018). In scientific research, observational studies (non-intervention) are designed for testing association/correlation between variables, while intervention studies, such as clinical trials, are for testing causal relations (Buhse et al., 2018). Misinterpreting correlations as causations can lead to serious consequences, such as wrong medical decisions (Buhse et al., 2018) or harmful misperception of certain groups of people in society (Richardson et al., 2014).

The misinterpretations are often attributed to humans' inherent tendency to conflate correlation and causation (Bleske-Rechek et al., 2015), and journalists' lack of scientific training or their pursuit of sensational effect (Fahnestock, 1998; Sumner et al., 2014). However, misinterpreting correlational findings as causal claims has been found not only in news stories and press releases written by journalists (Sumner et al., 2014), but also in research papers published by scientists. For example, Robinson et al. (2007) manually examined the methodologies of educational studies and found an increasing number of nonintervention studies using causal statements, from 34% in 1994 to 43% in 2004. Cofield et al. (2010), also through manual content analysis, found 31% of observational studies in obesity and nutrition inappropriately used causal language.

Scientists' inappropriate use of causal language may also be attributed to the pursuit of exaggeration in research contribution, or the lack of training in science writing, especially for non-English speaking researchers. For example, a number of linguistic studies have found that expressing and interpreting epistemic modality is not an easy task for learners of English as a second language (Holmes, 1982). There is also a lack of consistent

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4664–4674, Hong Kong, China, November 3–7, 2019.

teaching of strategies for thinking about probability and causality (Stanovich, 2010). With English as the most popular language for scientific writing, non-native English-speaking authors' native languages may also affect their choice of causal expressions (Hyland and Milton, 1997; Hyland, 1998; McEnery and Kifle, 2002; Kranich, 2009), adding to the challenge of establishing a shared standard of causal language use in the global science community.

In this study, we developed a prediction model to automatically identify correlational and causal statements in research conclusions. We then applied this model to the observational studies in PubMed to answer three research questions regarding causal language use. First, what is the overall ratio of causal lanugage use in observational studies? Second, what is the trend of causal language use? Third, do authors from different countries and language backgrounds use causal language at different levels? This study contributes to the field of computational social science by providing a new prediction model to identify causal language use in research findings, and by providing new evidence from large-scale analysis for answering research questions regarding causal language use in global science communication.

2 Related Work

2.1 Challenge in Describing Causal Relations

Making valid causal inferences from observational data is a challenging and risky process (Hernán and Robins, 2006). Linguistic expressions of causal relations vary greatly (Dunietz et al., 2017). Commonly used causal markers can be domain- and genre-dependent (Mulkar-Mehta et al., 2011). Therefore, choosing the appropriate linguistic cues to describe different causal levels has been identified as one of the main difficulties that student writers experience in managing epistemic meaning (Holmes, 1982). Though such difficulty is reflected in a number of best-practice guides for science writers (Zweig and DeVoto, 2018), suggestions in those guides are generally based on the judgment of only a few individuals (Adams et al., 2017), and there lacks consistent teaching of strategies for thinking about probability and causality (Stanovich, 1992). As a result, authors of observational studies sometimes wrongly extrapolate their results for clinical practice recommendations rather than suggesting follow-up randomized controlled trials (RCTs) (Prasad et al., 2013). Inappropriate use of causal language was then observed by researchers in different science reporting venues, such as leading research journals (e.g. Cofield et al., 2010; Brown et al., 2013; Lazarus et al., 2015), and press releases (e.g. Yavchitz et al., 2012).

2.2 Definition of Causal Relations

Detecting causal language use is closely related to the work of cause-effect relation extraction. NLP tasks such as SemEval-2007 (Girju et al., 2007) and SemEval-2010 (Hendrickx et al., 2009) included extracting cause-effect relations as a subtask of semantic relation extraction. Overall, the cause-effect relation could broadly refer to relations between events and entities (Bethard and Martin, 2008; Mirza and Tonelli, 2014; Chambers et al., 2014), clauses (Grivaz, 2010), and discourse arguments (Prasad et al., 2008).

Contrarily, some studies specifically focused on the causal relations expressed between independent variables and dependent variables. They normally formulate the task as a binary classification problem to distinguish causation from correlation (e.g. Cofield et al., 2010). To capture the different levels of certainty in research findings, Sumner et al. (2014) proposed a more fine-grained categorization with seven levels: *no mentioned relationship*, *statement of no relationship*, *statement of correlation*, *ambiguous statement of relationship*, *conditional statement of causation*, *statement of can*, and *statement of causation*.

In comparison, some defined patterns by the broader definition are not applicable to the narrower definition. For example, relations such as "purpose" and "motivation" (Dunietz et al., 2015) between news events are hard to be generalized to describe relations between variables in research findings. Meanwhile, discourse markers such as reporting verbs (Mihăilă et al., 2013) and conjunctions like "because" are commonly used indicators for causal relations at clause or discourse levels, but are rarely used to describe relations between variables. In this study, we adopted the narrow definition and focused on the causal relations expressed between dependent variable and independent variables.

2.3 Corpora for Causal Relation Analysis

Texts from different domains have been used for causal relation analysis, such as news articles (e.g.

Khoo et al., 1998; Girju et al., 2002; Girju, 2003; Prasad et al., 2008; Sumner et al., 2014), health and biomedical publications (e.g. Cofield et al., 2010; Khoo et al., 2000; Mihăilă et al., 2013; Sumner et al., 2014; Sharma et al., 2018), expository texts (Kaplan and Berry-Rogghe, 1991), and technical texts (Garcia et al., 1997). Mulkar-Mehta et al. (2011) compared the causal markers from different domains, some of which are domain-specific. Dunietz et al. (2017) also found that biomedical texts and other specialized fields may have context-specific language constructs about causality. A review of these prior studies showed that to date there is no existing corpus particularly focusing on research findings in the health domain.

2.4 Methods for Identifying Causal Relations

Both manual and machine learning approaches have been used to extract causal relations. Manual approaches often applied domain-specific, knowledge-based inferences (e.g. Kaplan and Berry-Rogghe, 1991) and hand-coded rules to encode the causal language cues (e.g. Garcia et al., 1997; Cofield et al., 2010; Lazarus et al., 2015; Chiu et al., 2017). However, such manual approaches often require an exhaustive list of linguistic cues, which costs a significant amount of human effort and its generalizability may be limited by the small sample size.

Meanwhile, machine learning algorithms, such as Logistic Regression (Bui et al., 2010), Naive Bayes (Chang and Choi, 2004), Decision Trees (Blanco et al., 2008), Conditional Random Fields (CRF) (Mihăilă and Ananiadou, 2013) and SVM, have also been used to extract causal relations. SVM was among the most frequently applied algorithms (e.g. Sarker and Gonzalez, 2015; Mirza and Tonelli, 2016).

Most recently, deep learning approaches have also been used to extract causal relations. For example, Miwa and Bansal (2016) proposed a LSTM-RNN model incorporated with rich linguistic structures to extract relations in nominal relation extraction in SemEval-2010 (Hendrickx et al., 2009), which outperformed the CNN-based models. Dasgupta et al. (2018) proposed a bidirectional LSTM model to extract causal relations from drug effect data and news articles, with better performance than the rule-based approaches and CRF.

Despite the progress on machine learning approaches for causal relation detection, due to the different definition of causal relations and the dataset, experimental results may not be necessarily comparable to each other. Their generalizability remains an open question for tasks focusing on research conclusions. To this end, Li et al. (2017) adopted a simplified taxonomy from (Sumner et al., 2014) for causal relations in research findings, and used SVM to develop a four-category classifier with a 0.718 F1-score, suggesting room for further performance improvement. In this study we will investigate traditional and deep learning methods for building causal language prediction model with a new, human-annotated corpus dedicated to research conclusions.

3 Corpus Construction

To perform automatic detection of causal language use in research findings, a data set consisting of both causal and non-causal statements is needed. We chose PubMed as the data source to construct the corpus. We focused on health because it is a research topic of high social impact, and inappropriate use of causal language in health literature has been reported in prior manual studies. PubMed provides convenient access to a large number of health research publications with rich metadata particularly useful for this study.

In this study we chose to analyze the abstracts instead of full-text articles because the PubMed abstracts are openly available to the public while many full-text articles are behind a paywall. The research papers in PubMed may include an unstructured abstract, or a structured abstract which usually has a conclusion subsection to present the major findings. For the convenience of locating research findings, we used the structured abstracts only.

We selected a sample of structured abstracts based on a three-dimensional stratified sampling strategy. First, to account for the vocabulary variation among different health issues, five common health topics —nutrition, diabetes, obesity, breast cancer, and cholesterol— were selected. Second, because different study designs can affect the strength of the research findings, and thus affect the language choice for correlational or causal relations, we sampled both observational studies and randomized controlled trials (RCTs) using PubMed's metadata property "Publication Type".

For observational studies, we also used an advanced keyword search query[1] to further identify four different types of observational studies (case-control, cross-sectional, retrospective cohort, and prospective cohort). Third, the abstracts were selected based on the length of the conclusion subsection. The XML files from PubMed contain occasional parsing errors, resulting in extremely long conclusions which actually included paragraphs in the following sections. For quality control purpose, we used the Stanford CoreNLP tool (Manning et al., 2014) to split the conclusion subsections into sentences and removed the articles with conclusions longer than four sentences. Eventually, articles were sampled with the conclusion length as 1, 2, 3, or 4 sentences.

We then constructed a coding schema to define the relation categories for research findings. Drawing on prior studies of factuality (Kilicoglu et al., 2015) and causal strength (Sumner et al., 2014), we defined a causal level taxonomy as a simplified combination of the taxonomies in the two studies: *correlational*, *conditional causal*, *direct causal*, and *no relationship*. Table 1 lists the category definitions and some common language cues used to identify the relation type for each category. Table 2 shows the examples of sentences with different relation types. Sometimes a sentence may contain language cues indicating a causal relationship, such as "a reliable way to determine"; however, if the sentence describes the function of certain tools or diagnoses rather than the explicit relations between independent variable and dependent variables, it should be labelled as "no relationship" (as shown in Example 6 of Table 2). Meanwhile, sentences that discuss study limitations, as shown in Example 7 of Table 2, should be labelled as "no relationship".

A sample of 30 abstracts were randomly selected for the inter-coder reliability test. Specifically, two annotators labelled the relation type for each sentence from the 30 conclusion subsections, and also highlighted the linguistic indicators. The overall Cohen's Kappa agreement (Cohen, 1960) was 0.98, indicating a near-perfect inter-coder agreement (McHugh, 2012). Disagreements in the annotation were later resolved by the two annotators through discussion.

One of the annotators, who has linguistics background, then annotated the rest of the data. The annotator also marked unsure cases and brought to team discussion before finalizing the annotations. The corpus contains 3,126 annotated sentences, of which 3,061 sentences contained only one type of relations, and the remaining 65 sentences had more than one type of relations. Table 3 shows the relation type distribution of single-class sentences in the corpus.

4 Experiments and Evaluation

We trained and evaluated four machine learning approaches for classifying causal language use in conclusion sentences. The learning approaches are bag-of-words based Linear SVM, Bidirectional Recurrent Neural Network, BERT (Devlin et al., 2018), and BioBERT (Lee et al., 2019).

4.1 Experimental Setup

Linear SVM. We used the Scikit Learn vectorizer with unigrams and bigrams as feature set to convert texts to sparse vectors. Adding tri-grams did not improve model performance. Performance on *count* and *tf-idf* vectorizations were similar. The parameter C in LinearSVM was set to 0.1 after tuning.

BiRNN. Different from the traditional bag-of-words methods that convert a text to a sparse vector, BiRNN and BERT are deep learning approaches designed to transform a text to a dense vector that can better represent the meaning of the text. In our implementation, we used the combination of two word embeddings: the PMC embedding based on full-text research publications from PubMed Central (PMC) [2], and the universal GloVe embedding. We found that the PMC embedding did better than GloVe, and their combination performed the best. In addition, we used GRU recurrent network units, which are generally simpler and faster than LSTM units.

BERT. BERT is the latest method for pre-training language representations, which has achieved the state-of-the-art results on a wide array of NLP tasks, including sentence classification. We used the Pytorch version of BERT with its pre-trained model (Cased BERT-Base, which performs better than the Uncased one) [3]. Because the distribution of sentence relationships in our corpus is unbalanced, we revised the loss function in the original

[1] Public Health: Search Strategies by Study Type (University of Adelaide)

[2] http://bio.nlplab.org/

[3] https://github.com/huggingface/pytorch-pretrained-BERT

Relation Type	Description	Language Cue
Correlational	The statement describes the association between variables, but causation cannot be explicitly stated.	association, associated with, predictor, at high risk of
Conditional Causal	The statement shows that one variable directly changes the other. However, the relation carries an element of doubt in it, which is normally via hedges or modalities.	increase, decrease, lead to, effect on, contribute to, result in *Cues indicating doubt:* may, might, appear to, probably
Direct Causal	The statement says that the independent variable directly alters the dependent variable.	increase, decrease, lead to, effective in, contribute to, reduce
No Relationship	No correlation or causation relationship is mentioned in the statement.	

Table 1: A taxonomy of relation types and examples of commonly used language cues.

Annotation	Example Sentence
Correlational	1 The findings from this large prospective study show that men who are taller and who have greater adiposity have an elevated risk of high-grade prostate cancer and prostate cancer death. 2 The association of high PCT and CRP was no more predictive of mortality than high CRP.
Conditional Causal	3 MTHFR A1298C polymorphism might contribute to an increased risk of breast cancer and/or ovarian cancer susceptibility.
Direct Causal	4 Participatory community-based nutrition education for caregivers improved child dietary diversity even in a food insecure area. 5 The adoption of a comprehensive preoperative multidisciplinary approach led to significant improvements in the postoperative outcomes and also in the compliance to the postoperative follow-up.
No Relationship	6 A quick foot-of-the-bed clinical assessment is not a reliable way to determine frailty. 7 This approach may, however, be difficult to implement on a large scale.

Table 2: Examples of sentences and annotated relation types.

Label	Count	Ratio
Correlational	998	32.6%
Conditional Causal	213	7.0%
Direct Causal	494	16.1%
No Relationship	1356	44.3%
Total	3061	100.0%

Table 3: Distribution of sentence types in the corpus.

code to ensure that the penalty on misclassifying rare classes was high.

BioBERT. BioBERT is a domain specific language representation model that is based on BERT and further pre-trained on large-scale biomedical corpora. BioBERT has shown to outperform BERT on three representative biomedical text mining tasks (Lee et al., 2019). In our experiment, we used the same BERT code as described above, and replaced the Cased BERT-Base pre-trained model with the BioBERT model.

4.2 Evaluation

To compare the performance of the above four learning approaches, we evaluated them on our annotated corpus via 5-fold stratified cross-validation. The result in Table 4 shows that BioBERT performed the best by all measures, followed by BERT, BiRNN and then LinearSVM. BioBERT achieved 0.881 macro-averaged F1 score with fairly balanced performance on each category except the conditional causal one, which only covers 7% data in the training set. See the performance scores per category in Table 5 and the confusion matrix in Fig. 1.

	Precision	Recall	Accuracy	F1
LinearSVM	0.739	0.711	0.772	0.722
BiRNN	0.801	0.825	0.836	0.811
BERT	0.867	0.883	0.889	0.874
BioBERT	0.878	0.886	0.901	0.881

Table 4: Model performance: BioBERT$>$ BERT $>$ BiRNN $>$ LinearSVM. (F1 represents macro-F1.)

4.3 Error Analysis

Error analysis of misclassified cases shows the most common disagreement between machine prediction and human annotation is between the categories of *direct causal* and *no relationship*, even though the two categories are not semanti-

	Precision	Recall	F1
Correlational	0.917	0.924	0.921
Conditional Causal	0.791	0.854	0.822
Direct Causal	0.878	0.858	0.868
No Relationship	0.915	0.906	0.911

Table 5: Detailed performance of BioBERT on each category of sentence relationship.

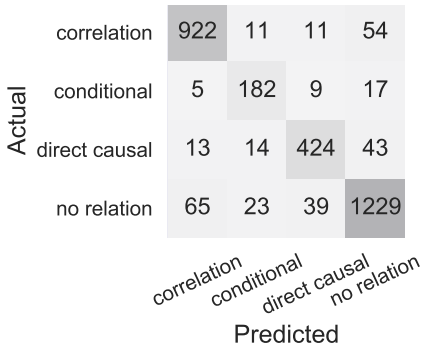

Figure 1: Confusion matrix. The prediction results on each test fold from 5-fold cross-validation were assembled together.

cally close in our dataset. This is largely caused by the confounding language cues in the *no relationship* examples. Specifically, a *no relationship* sentence that describes implications for future studies or functions of certain assessments may contain causal markers, such as "is critical to achieving", and "improved" in the following Examples 1 and 2, and thus mislead the prediction model. In addition, causal markers sometimes appeared in subordinate clauses instead of in the main clauses, which could also confuse the prediction model (see "the impact of" in Example 3). The prediction model also has some difficultly in recognizing the less commonly used causal markers, especially when nouns were used to describe causal relations, such as "cause" and "the key role of" in Examples 4 and 5.

Sentence examples:

1. However, training and experience of nurses in aromatherapy massage is critical to achieving positive results.
2. These findings pose the question: why has not the nutritional status of children improved, although the living conditions of their families have significantly improved?
3. Some studies that have assessed the impact of monitoring guidelines on clinical practice show only limited impact.
4. However, body mass index greater than 25 and smoking history are cause for caution.
5. Findings reinforce the key role of mvpa for childrens health.

4.4 Learning Curve

Our training dataset contains 3,061 sentences. To test whether the dataset size is sufficient to build a strong prediction model, we ran a learning curve test by setting aside 20% of the dataset for testing (i.e., about 600 examples), and then trained the BioBERT model with the 10%, 20%, ..., 100% of the remaining data (i.e., about 2500). We ran the experiment 5 times. As shown in Fig. 2, the average performance levels off around 60% of training data (i.e., about 1500), indicating that the training dataset is sufficient, although a larger dataset might help build a stronger model.

Figure 2: Learning curve. The average performance levels off around 60% of the training data, which is about 1500 examples.

5 Application to Observational Studies

In this section, we aim to investigate the causal language use in observational studies at a large scale. In 2014, the PubMed staff introduced *observational study* to their list of publication types[4]. To date 61,830 PubMed publications have been manually assigned to the observational study category. However, some complicated studies were assigned to multiple study design categories, such as both observational study and randomized controlled trial. After removing those with multiple study design types, we obtained 51,274 publications. Among them, 38,191 studies have structured abstracts with a conclusion subsection. For the reason mentioned in section 3, we do not consider those publications with more than 4 sentences in their conclusion subsections, since those with more than 4 sentences are often the result

[4]https://www.ncbi.nlm.nih.gov/mesh/68064888

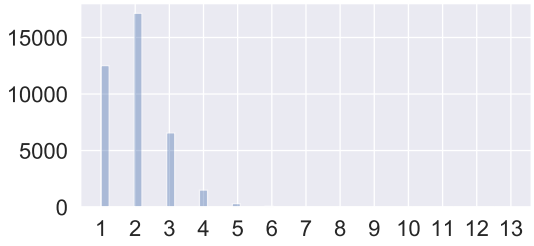

Figure 3: Distribution of number of sentences in the abstract conclusion subsections of the 38,191 observational studies. 98.8% of the publications have 4 or less sentences.

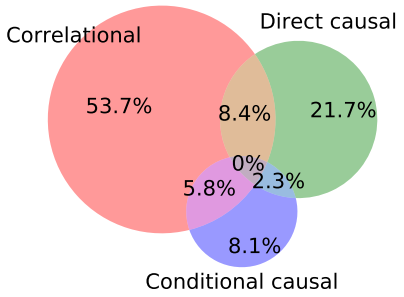

Figure 4: The constitution of the three relationships.

of PubMed metadata concatenating the conclusion subsection with the following paragraphs. Among the 38,191 publications, 98.8% have 4 or less sentences. (Fig. 3 shows the distribution of number of sentences included in the conclusion subsections.) The final data set of observational studies includes 37,746 publications, which in total contains 72,565 conclusion sentences.

We then applied the sentence classification model, as described in the previous section, to these conclusion sentences. The prediction results from the model are used to answer three research questions regarding causal language use in observational studies.

5.1 Results and Discussion

We aim to investigate the following questions:

RQ1. What is the overall ratio of causal lanugage use in observational studies?

The prediction result shows that among the 37,746 observational studies, 22% did not mention either correlation or causal relations. Because we would like to focus on distinguishing correlation vs. causal statements, the publications without mentioning relations were excluded, and the remaining 78% or 29,410 studies were used in further analyses.

Since PubMed also includes a small number of non-English publications with an English abstract, we used the PubMed metadata "language" to separate the 29,410 studies into the English subset (28,217 publications) and the non-English subset (1,193 publications).

The predictions were made on individual sentences. A conclusion subsection may contain multiple sentences; only 33% of the publications consist of one conclusion sentence.

Using the result from the 28,217 English publications, a Venn diagram (Fig. 4) was generated to show the proportion of studies mentioning one or more types of relations. It shows that 83.5% of studies contained only one type of relation in the conclusion: 53.7% correlational, 8.1% conditional causal, and 21.7% direct causal. Most of these studies contain only one sentence in the conclusion subsection. A manual examination of examples in the mixed categories found two main types of cases. The first type usually includes findings for multiple research questions. The second type uses multiple sentences to describe one finding; for example, the first sentence describes a correlation, and the second generalizes the correlation to a conditional or direct causal relation.

In this study we focus on the inappropriate use of direct causal language in observational studies. Therefore, we calculated the percentage of studies that used direct causal language in at least one finding sentence, and the percentage of studies that exclusively used direct causal language in their findings. Our result shows that 21.7% studies used direct causal exclusively in their conclusions, and 32.4% used some direct causal language. This result is close to the manual estimation reported in (Cofield et al., 2010), which is 31%.

RQ2. What is the trend of causal language use?

From now on, in the following figures and tables, for simplicity, we use word *causal* for the studies that used some direct causal language, and phrase *causal only* for those that used direct causal exclusively. Fig. 5 illustrates the ratios of causal language use and causal only from year 2013 to 2019. The two ratios follow the same pattern. Different from (Robinson et al., 2007), no obvious increase of causal language use was observed in our study.

One limitation of our study is that the PubMed labels of observational studies were introduced in 2014 (although some 2013 publications were an-

notated too); therefore, our dataset covers only seven years with partial data from 2013 and 2019. In the future we will design a classifier to automatically label observational studies, and then will be able to analyze the trend with more longitudinal data.

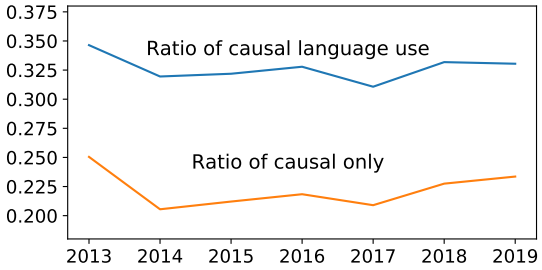

Figure 5: Causal language use over recent years.

RQ3. Do authors from different countries and language backgrounds use causal language at different levels? Prior studies in linguistics, as described in sections 1 and 2, have found that German, French, and Dutch translations tend to use stronger claims than the original English science publications. In this study, we will be able to compare causal language use in English and non-English publications, all with English abstracts. We hypothesize that if authors' native languages did affect their causal language use in English, the effect would be more pronounced in the English abstracts of the non-English publications, because these authors mainly used their native languages for science communication.

Language	Papers	Causal	Causal only
German	34	0.706	0.500
French	145	0.490	0.393
Chinese	80	0.487	0.388
Italian	32	0.469	0.375
Spanish	715	0.394	0.297
English	28217	0.323	0.217
Portugese	124	0.266	0.194

Table 6: Ratios of causal language use in studies published in English and other languages (sorted by *causal only*, reversely). Languages with fewer than 30 publications were excluded.

Table 6 shows the ratios of causal language use in the non-English subset. The result shows that publications in some languages, such as German, used much more causal language than those in English and Portugese.

Country	Papers	Causal	Causal only
Germany	909	0.421	0.295
Italy	1314	0.390	0.287
France	992	0.342	0.247
Spain	1621	0.365	0.239
UK	1071	0.331	0.222
Brazil	750	0.316	0.219
Japan	1299	0.296	0.216
China	919	0.301	0.215
Netherlands	797	0.359	0.213
Korea	772	0.284	0.201
US	4771	0.277	0.174

Table 7: Ratios of causal language use in different countries (sorted by *causal only*, reversely). Countries with fewer than 750 publications were excluded.

Table 7 shows the ratios in the English subset, broken down by the authors' countries. For each publication, each author's country was extracted from the PubMed affiliation metadata. Publications with authors from more than one country were excluded. The result shows that Germany has the highest ratio of causal language use (42.1%). In comparison, the US has the lowest ratio.

Because some languages are used in multiple countries and some countries did not publish many papers, not all results in Table 6 and Table 7 are comparable. However, we were able to identify four country-language pairs that correspond well to each other: Germany-German, Italy-Italian, France-French, and China-Chinese. In the case of other non-English languages, Portugese is majorly used by Brazil and Portugal, and Spanish is majorly used in Spain and Mexico. In the above four pairs, the ratios of causal language use in the non-English subsets were all higher than those in the English subset.

These results provide supportive evidence to the claim that causal language use varies across countries and languages, and thus challenge the notion of a shared consensus on causal language use in global science community.

6 Conclusion and Future Work

In this study we developed a manually-annotated corpus for causal statements in science publications, and used the corpus to train a prediction model to identify correlational, conditional causal, and direct causal statements in research conclu-

sions, achieving an accuracy of 0.90 and a macro-F1 score of 0.88. This prediction model provides an automated approach for identifying potential misuse of causal language in science writing. When applying the model to investigating association between causal language use and authors' countries or native languages, the result lends support to the claim that causal language use varies across countries and languages, and thus challenges the notion of a shared consensus of causal language use in the global science community. The result is also important for raising awareness towards recognizing disparities in causal language use in science communication.

In the future we will further our investigation by developing prediction models to automatically identify study designs in more fine-grained categories, such as cross-sectional, case-control, retrospective cohort, and prospective cohort. We will then be able to accurately identify inappropriate causal language use in weaker study design types, and also to compare among countries and languages in each specific study design types.

We also plan to extend our work to other domains in the future. In our current work we chose the biomedical domain as the starting point not only for the problem significance but also for the high-quality and rich metadata provided by PubMed that can help identify study types. By contrast, no such data are available in other domains yet. Toward this new goal, we also plan to develop a classification method to identify study types in other domains such as psychology and education.

The corpus and the code are available at https://github.com/junwang4/causal-language-use-in-science.

Acknowledgments

This work was partially supported by Syracuse University CUSE Grant Program and NVIDIA GPU Grant Program. We thank the reviewers for helpful comments, and Albert Wang for help with annotation.

References

Rachel C Adams, Petroc Sumner, Solveiga Vivian-Griffiths, Amy Barrington, Andrew Williams, Jacky Boivin, Christopher D Chambers, and Lewis Bott. 2017. How readers understand causal and correlational expressions used in news headlines. *Journal of experimental psychology: applied*, 23(1):1.

Steven Bethard and James H Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 177–180. Association for Computational Linguistics.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*.

April Bleske-Rechek, Katelyn M Morrison, and Luke D Heidtke. 2015. Causal inference from descriptions of experimental and non-experimental research: Public understanding of correlation-versus-causation. *The Journal of general psychology*, 142(1):48–70.

Isabelle Boutron and Philippe Ravaud. 2018. Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):2613–2619.

Andrew W Brown, Michelle M Bohan Brown, and David B Allison. 2013. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. *The American journal of clinical nutrition*, 98(5):1298–1308.

Susanne Buhse, Anne Christin Rahn, Merle Bock, and Ingrid Mühlhauser. 2018. Causal interpretation of correlational studies–analysis of medical news on the website of the official journal for german physicians. *PloS one*, 13(5):e0196833.

Quoc-Chinh Bui, Breanndán Ó Nualláin, Charles A Boucher, and Peter MA Sloot. 2010. Extracting causal relations on hiv drug resistance from literature. *BMC bioinformatics*, 11(1):101.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Du-Seong Chang and Key-Sun Choi. 2004. Causal relation extraction using cue phrase and lexical pair probabilities. In *International Conference on Natural Language Processing*, pages 61–70. Springer.

Kellia Chiu, Quinn Grundy, and Lisa Bero. 2017. spinin published biomedical literature: A methodological systematic review. *PLoS biology*, 15(9):e2002173.

Stacey S Cofield, Rachel V Corona, and David B Allison. 2010. Use of causal language in observational studies of obesity and nutrition. *Obesity facts*, 3(6):353–356.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316.

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*.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics*, 5:117–133.

Jeanne Fahnestock. 1998. Accommodating science: The rhetorical life of scientific facts. *Written communication*, 15(3):330–350.

Daniela Garcia et al. 1997. Coatis, an nlp system to locate expressions of actions connected by causality links. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 347–352. Springer.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics.

Roxana Girju, Dan I Moldovan, et al. 2002. Text mining for causal relations. In *FLAIRS conference*, pages 360–364.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18. Association for Computational Linguistics.

Cécile Grivaz. 2010. Human judgements on causation in french texts. In *LREC*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

Miguel A Hernán and James M Robins. 2006. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, pages 360–372.

Janet Holmes. 1982. Expressing doubt and certainty in english. *RELC journal*, 13(2):9–28.

Ken Hyland. 1998. *Hedging in scientific research articles*, volume 54. John Benjamins Publishing.

Ken Hyland and John Milton. 1997. Qualification and certainty in l1 and l2 students' writing. *Journal of second language writing*, 6(2):183–205.

Randy M Kaplan and Genevieve Berry-Rogghe. 1991. Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition*, 3(3):317–337.

Christopher SG Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 336–343. Association for Computational Linguistics.

Christopher SG Khoo, Jaklin Kornfilt, Robert N Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186.

Halil Kilicoglu, Graciela Rosemblat, Michael Cairelli, and Thomas Rindflesch. 2015. A compositional interpretation of biomedical event factuality. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 22–31.

Samantha Kleinberg and George Hripcsak. 2011. A review of causal inference for biomedical informatics. *Journal of biomedical informatics*, 44(6):1102–1112.

Svenja Kranich. 2009. Epistemic modality in english popular scientific texts and their german translations. *Trans-kom*, 2(1):26–41.

Clément Lazarus, Romana Haneef, Philippe Ravaud, and Isabelle Boutron. 2015. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC medical research methodology*, 15(1):85.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP Analysis of Exaggerated Claims in Science News. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Tony McEnery and Nazareth Amselom Kifle. 2002. Epistemic modality in argumentative essays of second-language writers. *Academic discourse*, pages 182–195.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Claudiu Mihăilă and Sophia Ananiadou. 2013. Recognising discourse causality triggers in the biomedical domain. *Journal of bioinformatics and computational biology*, 11(06):1343008.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC bioinformatics*, 14(1):2.

Paramita Mirza and Sara Tonelli. 2014. 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*, pages 2097–2106.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *The 26th international conference on computational linguistics*, pages 64–75. ACL.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*.

Rutu Mulkar-Mehta, Andrew S Gordon, Jerry R Hobbs, and Eduard Hovy. 2011. Causal markers across domains and genres of discourse. In *Proceedings of the sixth international conference on Knowledge capture*, pages 183–184. ACM.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Vinay Prasad, Joel Jorgenson, John PA Ioannidis, and Adam Cifu. 2013. Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. *Journal of clinical epidemiology*, 66(4):361–366.

Sarah S. Richardson, Cynthia R. Daniels, Matthew W Gillman, Janet Lynne Golden, Rebecca Kukla, Christopher W Kuzawa, and Janet E Rich-Edwards. 2014. Society: Don't blame the mothers. *Nature*, 512:131–132.

Daniel H Robinson, Joel R Levin, Greg D Thomas, Keenan A Pituch, and Sharon Vaughn. 2007. The incidence of causal statements in teaching-and-learning research journals. *American Educational Research Journal*, 44(2):400–413.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Raksha Sharma, Girish Palshikar, and Sachin Pawar. 2018. An unsupervised approach for cause-effect relation extraction from biomedical text. In *International Conference on Applications of Natural Language to Information Systems*, pages 419–427. Springer.

Keith E Stanovich. 1992. *How to think straight about psychology*. HarperCollins Publishers.

Keith E Stanovich. 2010. How to think straight about psychology, 9 th.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy Boy, and Christopher D Chambers. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, 349:g7015.

Amélie Yavchitz, Isabelle Boutron, Aida Bafeta, Ibrahim Marroun, Pierre Charles, Jean Mantz, and Philippe Ravaud. 2012. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS medicine*, 9(9):e1001308.

Mark Zweig and Emily DeVoto. 2018. Observational studies: Does the language fit the evidence? association vs. causation. http://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/. Accessed on 01/02/2018.