An NLP Analysis of Exaggerated Claims in Science News

Yingya Li
School of Information Studies
Syracuse University
yli48@syr.edu

Jieke Zhang
School of Information Studies
Syracuse University
jzhan150@syr.edu

Bei Yu
School of Information Studies
Syracuse University
byu@syr.edu

Abstract

The discrepancy between science and media has been affecting the effectiveness of science communication. Original findings from science publications may be distorted with altered claim strength when reported to the public, causing misinformation spread. This study conducts an NLP analysis of exaggerated claims in science news, and then constructed prediction models for identifying claim strength levels in science reporting. The results demonstrate different writing styles journal articles and news/press releases use for reporting scientific findings. Preliminary prediction models reached promising result with room for further improvement.

1 Introduction

On April 18, 2017 many science news agencies reported a new study on peer effects in health behavior (Aral and Nicolaides, 2017). Here are a few examples of the headlines:

AAAS: "Exercise is contagious, especially if you are a men"
MIT Sloan press release: "Turns out exercise is contagious"
Medscape: "Exercise may be contagious"
Gulfnews: "Exercise can be contagious, new social network analysis finds".

Regardless of the original finding, these news headlines interpreted and thus reported the finding with different levels of strength (using different verbs such as "*is*", "*may*", and "*can*").

This example illustrates a prominent problem in science communication that original scientific findings might be altered or distorted during the information spread process. Different information subsidies such as the university press and news releases have been widely used to deliver research findings. However, possibly caused by different writing purposes of scientists and journalists, those paraphrased versions of the original findings in the reporting may not be as accurate. For example, the university press release has been found to be a major source of misinformation (Sumner et al., 2014).

The ways in which information is framed along with how the audiences decode it has powerful impacts on public behaviors. Hence the aforementioned misinformation diffusion can cause misunderstanding of science findings. A possible approach for curbing such misinformation diffusion in science communication is to compare relevant findings reported in science news and the original journal articles, identifying the strength levels of their claims, and thus to warn writers and readers of potential exaggerations in the science reporting.

Such approach requires two steps: "claims pairing" and "claim strength identification". In this paper we focus on the second task, and leave the first task to future work. We explored the statement of causality in health-related science communication covered by academic journals, university press releases and news stories. We analyzed how causal triggers (i.e., verbs or verb phrases that express causal relations in claims) are associated with different levels of casual relations, using the open-dataset released by Sumner et al. (2014). Also, we developed text classification models to predict the strength levels of claims in academic papers and news articles.

This study seeks answers to the following research questions: (1) What are the linguistic factors distinguishing different reporting styles of journal articles and news/press releases? (2) What are the causal triggers for different levels of claim strength? (3) Is it feasible to automatically identify the strength of claims in science reporting and news? If so, what are the current achievement and challenges?

2 Related Work

In the NLP field scholars have tried to identify misinformation from different perspectives, including credibility prediction (Castillo et al., 2013), rumor detection (Qazvinian et al., 2011; Zubiaga et al., 2016) etc. Although satisfactory accuracy for automatic misinformation detection could be made, the effectiveness of discrediting misinformation on people's belief and perception remains unknown. Prior studies found that false information with exaggerated claims is designed to meet emotional needs and often emerges in situations of uncertainty (Silverman, 2015). For people with strong fixed views, encountering contradicting claims and arguments can cause them to strengthen their original belief. One possible way to reduce the continuing of misinformation is to explain why the information or myth is wrong by showing the rhetorical techniques such as the specific exaggeration that was used (Cook and Lewandowsky, 2012).

A relevant task of analyzing such rhetorical manipulation in science communication is to identify the strength of claims. Light et al. (2004) built a classifier to predict the levels of speculative language in sentences from biomedical abstracts. Vlachos and Craven (2010) also developed a classifier to detect the information certainty in biomedical text, using syntactic dependencies and logistic regression. Blake (2010) proposed a claim framework that tries to capture the ways an author communicates a scientific claim. The framework is built on the certainty of causal relations that were presented, which is closely related to strength identification.

The problem of identifying claim strength is also closely linked to several other science communication and science news reporting problems, especially on casual relation and exaggeration detection. Though many efforts have been made to analyze causal relations in claims (e.g., Mihaila et al., 2013; Sumner et al., 2014; Khoo et al., 2000), massive diffusion of unverified rumors fosters confusions about causation that could adversatively impact the public beliefs and decisions. Under this circumstance, readers' knowledge and personal judgment for claims of different issues will be challenged greatly.

Unlike previous studies mainly focusing on single domain, in the current work we studied claim strength across multiple domains/genres of both academic publication and news/press releases. We tried to automatically identify claim strength in science reporting with special focus on the different types of causal relationship. It is also the first step towards automatic identification of exaggeration and emotion manipulation in science news.

3 Experiment

3.1 Data

In this study, we use an open data set (*http://dx.doi.org/10.6084/m9.figshare.903704*) developed by Sumner et al. (2014). This corpus includes a sample of health-related journal articles and their corresponding press releases and news articles. After manually coding the strength levels of the main claims from the three sources, they found that the press release is a major source of exaggeration in science news reporting.

This open data set includes 462 health-related press releases and their corresponding claims in 668 associated journal articles and news stories. The primary causal claims in the journal articles, press releases and news reports are coded into seven categories with increasing strength of relationship: no mentioned relationship (Category "0"); statement of no relationship (Category "1"), statement of correlation (Category "2"), ambiguous statement of relationship (Category "3"), conditional statement of causation (Category "4"), statement of "can" (Category "5"), and statement of causation (Category "6"). Table 1 lists the category definitions and an example for each category.

3.2 Data Preprocessing

Adjusting category granularity: The original data set contains 1727 claims in 7 categories ("0"–"6"), with Category 6 (statement of causation) and Category 2 (statement of correlation) as the largest groups, accounting for 49% and 21% respectively. The other categories are relatively smaller.

To create a more balanced data set, we adjusted the category granularity, reducing the number of categories from 7 to 4. Category 0 was removed because it contains only 2 examples. Category 1 ("no relationship") remains the same. Category 3 is semantically close to Category 2, and thus was

Category	Statement	Category
No relationship mentioned – No relationship is mentioned	...we report the discovery and characterization of a unique core genome-encoded superantigen, providing new insights into the evolution of pathogenic S. aureus…	0
Statement of no relationship – Explicitly stating there is no relationship	…caesarean section by clinical officers does not result in a significant increase in maternal or perinatal mortality significant increase.	1
Statements of correlation – The IV and DV are associated, but causation cannot be explicitly stated	We found a strong graded relationship between increasing levels of psychological distress and the likelihood of being awarded a new disability pension.	2
Ambiguous statement of relationship – It is unclear what the strength of relationship of these statement is. The statement could mean that IV causes DV, or that the two variables are associated – either would be applicable.	…high levels of a protein called SGK1 are linked with infertility, while low levels of it make a woman more likely to have a miscarriage…	3
Conditional statement of causation – Causal statements show that the IV directly changes the DV. Conditional causal statements carry an element of doubt in them.	Genetic-screening trial could reduce drug side-effects.	4
Statement of "can" - The word "can" is unique as a statement of relationship in that it implies that the IV always has the potential to directly change the DV. it is a stronger statement than any conditional statement of causation.	Chocolate every day can reduce risk of heart disease.	5
Statements of causation – The strongest statements are statements of causation. This statement says that the IV directly alters the DV.	…three antiviral agents we studied significantly reduced the levels of Ab and P-tau…	6

IV: independent variable, DV: dependent variable

Table 1: Examples of different type of causal claims based on their strength.

merged into Category 2 ("correlation"). Categories 4 and 5 were merged into new Category 4 ("conditional causation") because both are weaker levels of causal relationships. Liberman (2011) found that although biomedical scientists clearly distinguished "may cause" (Category 4) and "can cause" (Category 5) types of relationships, science journalists seem not to distinguish them anyway. Category 6 ("causation") remains unchanged as the definitive statement of causation.

After adjusting the claim strength granularity, the original data was converted to four main categories: "no relationship" (Category 1), "correlation" (Category 2), "conditional causation" (Category 4), and "causation" (Category 6). Table 2 shows the distribution of each category before and after merging in the open dataset.

Separating training and testing data: The original data set contains 462 spreadsheets, one for each press release. Each spreadsheet documented the science claims reported in the original journal articles, and their paraphrased versions in the press releases and various news articles. Since all claims in the same spreadsheet involve the same science topic, we kept all statements from the same spreadsheet altogether either in training or in testing set to ensure the generalizability of the trained classifier. Specifically, statements from the first 300 spreadsheets were used for training and the rest 162 for testing.

Separating statements from journals and news/press: An important feature in academic writing is cautions language, often called "hedging" or "vague language", which may differ from the writing style in journalism. To test the homogeneity in writing style, we examined the hedging words in the training data using the Bioscope corpus (Szarvas et al., 2008). The Bioscope corpus marked a number of hedging cues in the abstracts of research articles, such as "may", "suggest", "indicate that", "whether", "appears". It is the most comprehensive hedging cues collection for biomedical writings we can find so far.

We calculated the document percentage of the statements with hedging words in the training data and consistently higher occurrences in journals than in news/press articles among all categories. See Table 3 for the distribution. The difference is the highest in Category 1 ("no relationship"), where hedges occurred in 81.5% journal claims but only in 58.6% press/news claims.

Due to the difference in writing style, we further separated statements in journal articles from those in press/news reports, and prepared training and testing data sets for each genre. Table 4 shows the distribution of statements after training/testing and journal/press separation.

Even though researchers claimed publication and reporting bias against negative findings (Dwan et al, 2008), our data consist of paired statements from different reporting sources; the percentage of the biased reporting should be comparable in journal articles and press releases. However, the category distribution in Table 4 shows that in journal articles a lot more correlations are reported, while in news/press releases more causation relations are reported. This observation along with the hedging words distribution supports our argument for the genre difference be between journal articles and news/press releases, justifying our decision to separate the statements according to their sources.

We did not further separate press release and news article to avoid overly small data sets, assuming no significant style difference in these two genres.

Claim Strength	**Original**	**Merged**
1 (no relationship)	82	82
2 (correlation)	366	519
3 (ambiguous relation)	153	
4 (conditional causation)	163	278
5 (statements of "can")	115	
6 (causation)	846	846
Total	1725	1725

Table 2: Claim strength distribution in the open dataset before and after category adjusting.

Claim Strength	**Journal Count (Percentage)**	**News/press Count (Percentage)**
1 (no relationship)	.815	.586
2 (correlation)	.756	.698
4 (conditional causation)	1.00	.984
6 (causation)	.706	.582
Total	.759	.690

Table 3: Hedging words distribution in journal and news/press.

Claim Strength	**Journal Train**	**Journal Test**	**News/ Press Train**	**News/ Press Test**
1	27 (.050)	11 (.039)	29 (.048)	15 (.050)
2	213 (.397)	115 (.405)	126 (.208)	65 (.218)
4	51 (.095)	24 (.085)	127 (.209)	76 (.255)
6	245 (.457)	134 (.472)	325 (.535)	142 (.477)
Total	536	284	607	298

Notes: numbers in the brackets are the percentages.

Table 4: Statement distribution after source and training/testing separation.

3.3 Feature Extraction

We constructed four feature vectors using different representations: 1) BOW: simple bag-of-words; 2) B-BOW: bag-of-words with the bolded linguistic cues that are manually-highlighted in the open data set; 3) N-BOW: bag-of-words with doubled negation words in the statements; 4) E-BOW: bag-of-words enriched with enhanced dependency parsing. We did not do stemming in order to keep word inflections. We did not remove stopwords because function words are likely style markers, for example "*that*" could indicate a subordinate clause.

For 2), the bolded linguistic cues (e.g., "*associated*", "*increased risk*", "*appears to offer*") were words/phrases labeled by annotators for identifying the claim strength. For 3), we searched for all the negation words (e.g., "*no*", "*not*") marked in the Bioscope corpus, and then doubled their occurrences in the statements by appending these words to the end of that statement (e.g. "*Water softeners provided no additional benefit to usual care.*" becomes "*Water softeners provided no additional benefit to usual care. no*"). For 4), we used the Stanford dependency parsing to extract all enhanced dependency relations in the statement. Dependency labels like *nsubj* and dependency words are tokenized separately and used as word features alone or combined with BOW to train our model).

For example,

Original statement from the open data set:

"A quick and cheap test could save the lives of babies born with congenital heart defects. (Category 4)"

Dependency words:

"test- A- test- quick- quick- and- quick- cheap- save- test- save- could- ROOT- save- lives- the- save- lives-

Claim Strength	MNB (tf)		SVM (boolean)		SVM (tf)	
	Journal	Press	Journal	Press	Journal	Press
1 (no relationship)	.632	.000	.696	.261	.667	.190
2 (correlation)	.649	.512	.629	.537	.639	.512
4 (conditional causation)	.400	.759	.766	.847	.783	.833
6 (causation)	.670	.748	.709	.784	.716	.768
Macro-average F1 score	.587	.505	.700	**.607**	**.716**	.576

Table 5: Classification accuracy of BOW unigram approach.

Claim Strength	Journal (SVM-tf)	Press (SVM-boolean)
1 (no relationship)	.667	.273
2 (correlation)	.648	.508
4 (conditional causation)	.826	.825
6 (causation)	.730	.780
Macro-average F1 score	**.718**	.596

Table 6: Classification accuracy of BOW unigram+bigram approach (using the best unigram model).

Claim Strength	B-BOW		N-BOW		E-BOW	
	Journal	Press	Journal	Press	Journal	Press
1 (no relationship)	.522	.182	.526	.105	.696	.250
2 (correlation)	.642	.542	.636	.512	.626	.545
4 (conditional causation)	.727	.821	.766	.836	.766	.831
6 (causation)	.702	.780	.716	.761	.704	.770
Macro-average F1 score	.648	.581	.661	.554	.698	.599

Table 7: Classification accuracy of B-BOW, N-BOW, and E-BOW approach.

lives- of- of- babies- babies- born- born- with- defects- congenital- defects- heart- with- defects"
Dependency tags:
"det amod cc conj nsubj aux root det dobj prep pobj vmod prep amod nn pobj"

The final vector is a combination of the three parts above.

3.4 Classification Results

Unigram features: We built two unigram models using Multinomial Naïve Bayes (MNB) and SVMs (Liblinear) with default settings in the Scikit Learn toolkit. Macro F1 scores are reported for evaluating the model performance in Table 5. For journal articles, SVM (with term frequency) has the best performance (F1 score = .716). For press/news articles SVM (with Boolean vectors) performed the best (F1 score = .607). Both models performed significantly better than the random guess baseline .25. Overall the model for the Journal genre performs better than the model for the press/news genre. Category wise the "no relationship" category has the lowest F1 scores, especially for statements in news/press releases. The "conditional causation" category has the highest F1 score among all claim strengths.

Enriched features: We continued to use SVM to build more models with enrich features. Adding bigrams resulted in slightly higher F1 score (.718) for journal and lower F1 score (.596) for press (as shown in Table 6). Therefore, we kept using the unigram features in later experiments. Table 7 reports the best classification results for the rest of each representation method mentioned in Section 3.3. As for B-BOW, we trained our model with bolded words only (with term frequency), bolded words only (with Boolean), and bolded words combined with the original statements (with term frequency).

4 Error Analysis

Error analysis shows that the classifier has a lot more to learn, such as variations in negation and distracting relationships mentioned in subordinate clauses. Analysis on the error cases in both journal articles and press releases shows that the most common disagreement is between categories 2 and 6, even though the two categories are not semantically close in the open dataset. This is largely caused by the location of the causal triggers for claim strength.

To further test the difficulty of identifying these two types of claim strength, we extracted about 50 statements in categories 2 and 6 from our misclassified cases and then invited two graduate students to judge their strength. The F1 scores compared to the ground truth (labeled score) were .440 and .630, with many Category 6 misjudged into Category 2 and vice versa, which is consistent to the machine performance. This low human performance also suggests the challenge of correctly identifying the claim strength even for well-educated readers.

5 Conclusion

In this study, we conducted an NLP analysis of claim strength and constructed prediction models for identifying claim strength levels in science reporting. Our best models reached .718 F1 score for distinguishing claim strengths in journal articles, and .607 F1 score in news/press releases, with very high performance for identifying conditional causations. Our analysis shows even though scientific writing follows a well-defined style, scientists' and journalists' creative use of language still poses significant challenge to our task. The major challenges are the variations in negation and distracting relationships mentioned in subordinate clauses for correlation and causation statements. We will conduct deeper syntactic analysis to improve the model performance in our future work.

References

Sinan Aral and Christos Nicolaides. 2017. Exercise contagion in a global social network. *Nature Communications* 8.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of biomedical informatics* 43(2):173–189

Kerry Dwan, Douglas G Altman, Juan A Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J Easterbrook, Erik Von Elm, Carrol Gamble, et al. 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS one* 3(8): e3081.

John Cook and Stephan Lewandowsky. 2011. *The debunking handbook*. Sevloid Art.

Panagiotis Takis Metaxas Eni Mustafaraj Markus Strohmaier Harald Schoen Gayo-Avello, Daniel Peter Gloor, Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Research* 23(5):560–588.

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*. Association for Computational Linguistics, pages 336–343.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*. Association for Computational Linguistics, pages 17– 24.

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

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Compu- tational Linguistics, pages 1589–1599.

Craig Silverman. 2015. Lies, damn lies, and viral content. how news websites spread (and debunk) online rumors, unverified claims, and misinformation. *Tow Center for Digital Journalism.*

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aime ́e Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective ob servational study. *Bmj* 349:g7015.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, pages 38–45.

Andreas Vlachos and Mark Craven. 2010. Detecting speculative language using syntactic dependencies and logistic regression. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*. Association for Computational Linguistics, pages 18–25.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11(3): e0150989.