Classifying Negative Findings in Biomedical Publications

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

Abstract

Publication bias refers to the phenomenon that statistically significant, "positive" results are more likely to be published than non-significant, "negative" results. Currently, researchers have to manually identify negative results in a large number of publications in order to examine publication biases. This paper proposes an NLP approach for automatically classifying negated sentences in biomedical abstracts as either reporting negative findings or not. Using multinomial naïve Bayes algorithm and bag-of-words features enriched by parts-of-speeches and constituents, we built a classifier that reached 84% accuracy based on 5-fold cross validation on a balanced data set.

1 Introduction

Publication bias refers to the phenomenon that statistically significant, "positive" results are more likely to be published than non-significant, "negative" results (Estabrook et al., 1991). Due to the "file-drawer" effect (Rosenthal, 1979), negative results are more likely to be "filed away" privately than to be published publicly.

Publication bias poses challenge for an accurate review of current research progress. It threatens the quality of meta-analyses and systematic reviews that rely on published research results (e.g., the Cochrane Review). Publication bias may be further spread through citation network, and amplified by citation bias, a phenomenon that positive results are more likely to be cited than negative results (Greenberg, 2009).

To address the publication bias problem, some new journals were launched and dedicated to publishing negative results, such as the Journal of Negative Results in Biomedicine, Journal of Pharmaceutical Negative Results, Journal of Negative Results in Ecology and Evolutionary Biology, and All Results Journals: Chem. Some quantitative methods like the funnel plot (Egger et al., 1997) were used to measure publication bias in publications retrieved for a certain topic.

A key step in such manual analysis is to examine the article abstracts or full-texts to see whether the findings are negative or not. For example, Hebert et al. (2002) examined the full text of 1,038 biomedical articles whose primary outcomes were hypothesis testing results, and found 234 (23%) negative articles. Apparently, such manual analysis approach is time consuming. An accurate, automated classifier would be ideal to actively track positive and negative publications.

This paper proposes an NLP approach for automatically identifying negative results in biomedical abstracts. Because one publication may have multiple findings, we currently focus on classifying negative findings at sentence level: for a sentence that contains the negation cues "no" and/or "not", we predict whether the sentence reported negative finding or not. We constructed a training data set using manual annotation and convenience samples. Two widely-used text classification algorithms, Multinomial naive Bayes (MNB) and Support Vector Machines (SVM), were compared in this study. A few text representation approaches were also compared by their effectiveness in building the classifier. The approaches include (1) bag-of-words (BOW), (2) BOW with PoS tagging and shallow parsing, and (3) local contexts of the negation cues "no" and 'not", including the words, PoS tags, and constituents. The best classifier was built using MNB and bag-of-words features enriched with PoS tags and constituent markers. The best performance is 84% accuracy based on 5-fold cross validation on a balanced data set.

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 19–23,
Baltimore, Maryland USA, June 26-27 2014.

2 Related work

The problem of identifying negative results is related to several other BioNLP problems, especially on negation and scientific claim identification.

The first relevant task is to identify negation signals and their scopes (e.g., Morante and Daelemans, 2008;2009; Farkas et al., 2010; Agarwal et al., 2011). Manually-annotated corpora like BioScope (Szarvas et al., 2008) were created to annotate negations and their scopes in biomedical abstracts in support of automated identification. This task targets a wide range of negation types, such as the presence or absence of clinical observations in narrative clinical reports (Chapman et al., 2001). In comparison, our task focuses on identifying negative findings only. Although not all negations report negative results, negation signals are important rhetorical device for authors to make negative claims. Therefore, in this study we also examine precision and recall of using negation signals as predictors of negative findings.

The second relevant task is to identify the strength and types of scientific claims. Light et al. (2004) developed a classifier to predict the level of speculations in sentences in biomedical abstracts. Blake (2010) proposed a "claim framework" that differentiates explicit claims, observations, correlations, comparisons, and implicit claims, based on the certainty of the causal relationship that was presented. Blake also found that abstracts contained only 7.84% of all scientific claims, indicating the need for full-text analysis. Currently, our preliminary study examines abstracts only, assuming the most important findings are reported there. We also focus on coarse-grained classification of positive vs. negative findings at this stage, and leave for future work the task of differentiating negative claims in finer-granularity.

3 The NLP approach

3.1 The definition of negative results

When deciding what kinds of results count as "negative", some prior studies used "non-significant" results as an equivalent for "negative results" (e.g. Hebert et al., 2002; Fanelli, 2012). However, in practice, the definition of "negative results" is actually broader. For example, the Journal of Negative Results in Biomedicine (JNRBM), launched in 2002, was devoted to publishing "unexpected, controversial, provocative and/or negative results," according to the journal's website. This broader definition has its pros and cons. The added ambiguity poses challenge for manual and automated identification. At the same time, the broader definition allows the inclusion of descriptive studies, such as the first JNRBM article (Hebert et al., 2002).

Interestingly, Hebert et al. (2002) defined "negative results" as "non-significant outcomes" and drew a negative conclusion that "prominent medical journals often provide insufficient information to assess the validity of studies with negative results", based on descriptive statistics, not hypothesis testing. This finding would not be counted as "negative" unless the broader definition is adopted.

In our study, we utilized the JNRBM articles as a convenience sample of negative results, and thus inherit its broader definition.

3.2 The effectiveness of negation cues as predictors

The Bioscope corpus marked a number of negation cues in the abstracts of research articles, such as "not", "no", "without", etc. It is so far the most comprehensive negation cue collection we can find for biomedical publications. However, some challenges arise when applying these negation cues to the task of identifying negative results.

First, instead of focusing on negative results, the Bioscope corpus was annotated with cues expressing general negation and speculations. Consequently, some negation cues such as "unlikely" was annotated as a speculation cue, not a negation cue, although "unlikely" was used to report negative results like

> "*These data indicate that changes in Wnt expression per se are* ***unlikely*** *to be the cause of the observed dysregulation of β-catenin expression in DD*" (PMC1564412).

Therefore, the Bioscope negation cues may not have captured all negation cues for reporting negative findings. To test this hypothesis, we used the JNRBM abstracts (N=90) as a convenience sample of negative results, and found that 81 abstracts (88.9%) contain at least one Bioscope negation cue. Note that because the JNRBM abstracts consist of multiple subsections "background", "objective", "method", "result", and "conclusion", we used the "result" and "conclusions" subsections only to narrow down the search range.

Among the 9 missed abstracts, 5 used cues not captured in Bioscope negation cues: "insufficient", "unlikely", "setbacks", "worsening", and "underestimates". However, the authors' writing style might be affected by the fact that JNRBM is dedicated to negative results. One hypothesis is that the authors would feel less pressure to use negative tones, and thus used more variety of negation words. Hence we leave it as an open question whether the new-found negation cues and their synonyms are generalizable to other biomedical journal articles.

The rest 4 abstracts (PMC 1863432, 1865554, 1839113, and 2746800) did not report explicit negation results, indicating that sometimes abstracts alone are not enough to decide whether negative results were reported, although the percentage is relatively low (4.4%). Hence, we decided that missing target is not a major concern for our task, and thus would classify a research finding as positive if no negation cues were found.

Second, some positive research results may be mistaken as negative just because they used negation cues. For example, "without" is marked as a negation cue in Bioscope, but it can be used in many contexts that do not indicate negative results, such as

> "*The effects are consistent with or without the presence of hypertension and other comorbidities and across a range of drug classes.*" (PMC2659734)

To measure the percentage of false alarm, we applied the aforementioned trivial classifier to a corpus of 600 abstracts in 4 biomedical disciplines, which were manually annotated by Fanelli (2012). This corpus will be referred to as "Corpus-600" hereafter. Each abstract is marked as "positive", "negative", "partially positive", or "n/a", based on hypothesis testing results. The latter two types were excluded in our study. The trivial classifier predicted an abstract as "positive" if no negation cues were found. Table 1 reported the prediction results, including the precision and recall in identifying negative results. This result corroborates with our previous finding that the inclusiveness of negation cues is not the major problem since high recalls have been observed in both experiments. However, the low precision is the major problem in that the false negative predictions are far more than the true negative predictions. Hence, weeding out the negations that did not report negative results became the main purpose of this preliminary study.

Discipline	#abstracts	Precision	Recall
Psychiatry	140	.11	.92
Clinical Medicine	127	.16	.94
Neuroscience	144	.20	.95
Immunology	140	.18	.95
Total	551	.16	.94

Table 1: results of cue-based trivial classifier

3.3 Classification task definition

This preliminary study focuses on separating negations that reported negative results and those not. We limit our study to abstracts at this time. Because a paper may report multiple findings, we performed the prediction at sentence level, and leave for future work the task of aggregating sentence-level predictions to abstract-level or article-level. By this definition, we will classify each sentence as reporting negative finding or not. A sentence that includes mixed findings will be categorized as reporting negative finding.

"Not" and "no" are the most frequent negation cues in the Bioscope corpus, accounting for more than 85% of all occurrences of negation cues. In this study we also examined whether local context, such as the words, parts-of-speeches, and constituents surrounding the negation cues, would be useful for predicting negative findings. Considering that different negation cues may be used in different contexts to report negative findings, we built a classifier based on the local contexts of "no" and "not". Contexts for other negation cues will be studied in the future.

Therefore, our goal is to extract sentences containing "no" or "not" from abstracts, and predict whether they report negative findings or not.

3.4 Training data

We obtained a set of "positive examples", which are negative-finding sentences, and a set of "negative examples" that did not report negative findings. The examples were obtained in the following way.

Positive examples. These are sentences that used "no" or "not" to report negative findings. We extracted all sentences that contain "no" or "not" in JNRBM abstracts, and manually marked each sentence as reporting negative findings or

not. Finally we obtained 158 sentences reporting negative findings.

To increase the number of negative-finding examples and add variety to writing styles, we repeat the above annotations to all Lancet abstracts ("result" and "finding" subsections only) in the PubMed Central open access subset, and obtained 55 more such sentences. Now we have obtained 213 negative-finding examples in total.

Negative examples. To reduce the workload for manual labeling, we utilized the heuristic rule that a "no" or "not" does not report negative result if it occurs in a positive abstract, therefore we extracted such sentences from positive abstracts in "Corpus-600". These are the negative examples we will use. To balance the number of positive and negative examples, we used a total of 231 negative examples in two domains (132 in clinical medicine and 99 in neuroscience) instead of all four domains, because there are not enough positive examples.

Now the training data is ready for use.

3.5 Feature extraction

We compared three text representation methods by their effectiveness in building the classifier. The approaches are (1) BOW: simple bag-of-words, (2) E-BOW: bag-of-words enriched with PoS tagging and shallow parsing, and (3) LCE-BOW: local contexts of the negation cues "no" and 'not", including the words, PoS tags, and constituents. For (2) and (3), we ran the OpenNLP chunker through all sentences in the training data. For (3), we extracted the following features for each sentence:

- The type of chunk (constituent) where "no/not" is in (e.g. verb phrase "VP");
- The types of two chunks before and after the chunk where "not" is in;
- All words or punctuations in these chunks;
- The parts-of-speech of all these words.

See Table 2 below for an example of negative finding: row 1 is the original sentence; row 2 is the chunked sentence, and row 3 is the extracted local context of the negation cue "not". These three representations were then converted to feature vectors using the "bag-of-words" representation. To reduce vocabulary size, we removed words that occurred only once.

(1)	Vascular mortality did not differ significantly (0.19% vs 0.19% per year, p=0.7).
(2)	"[NP Vascular/JJ mortality/NN] [VP did/VBD not/RB differ/VB] [ADVP significantly/RB] [PP (/-LRB-] [NP 019/CD %/NN] [PP vs/IN] [NP 019/CD %/NN] [PP per/IN] [NP year/NN] ,/, [NP p=07/NNS] [VP)/-RRB-] ./."
(3)	"na na VP ADVP PP did not differ significantly VBD RB VB RB"

Table 2: text representations

3.6 Classification result

We applied two supervised learning algorithms, multinomial naïve Bayes (MNB), and Support Vector Machines (Liblinear) to the unigram feature vectors. We used the Sci-kit Learn toolkit to carry out the experiment, and compared the algorithms' performance using 5-fold cross validation. All algorithms were set to the default parameter setting.

Representation		MNB	SVM
Presence vs. absence	BOW	.82	.79
	E-BOW	.82	.79
	LCE-BOW	.72	.72
tf	BOW	.82	.79
	E-BOW	.84	.79
	LCE-BOW	.72	.72
Tfidf	BOW	.82	.75
	E-BOW	**.84**	.73
	LCE-BOW	.72	.75

Table 3: classification accuracy

Table 3 reports the classification accuracy. Because the data set contains 213 positive and 231 negative examples, the majority vote baseline is .52. Both algorithms combined with any text representation methods outperformed the majority baseline significantly. Among them the best classifier is a MNB classifier based on enriched bag-of-words representation and tfidf weighting. Although LCE-BOW reached as high as .75 accuracy using SVM and tfidf weighting, it did not perform as well as the other text representation methods, indicating that the local context with +/- 2 window did not capture all relevant indicators for negative findings.

Tuning the regularization parameter C in SVM did not improve the accuracy. Adding bi-

grams to the feature set resulted in slightly lower accuracy.

4 Conclusion

In this study we aimed for building a classifier to predict whether a sentence containing the words "no" or "not" reported negative findings. Built with MNB algorithms and enriched bag-of-words features with tfidf weighting, the best classifier reached .84 accuracy on a balanced data set.

This preliminary study shows promising results for automatically identifying negative findings for the purpose of tracking publication bias. To reach this goal, we will have to aggregate the sentence-level predictions on individual findings to abstract- or article-level negative results. The aggregation strategy is dependent on the decision of which finding is the primary outcome when multiple findings are present. We leave this as our future work.

Reference

S. Agarwal, H. Yu, and I. Kohane, I. 2011. BioNOT: A searchable database of biomedical negated sentences. BMC bioinformatics, 12: 420.

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

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. 2001. Evaluation of negation phrases in narrative clinical reports. *Proceedings of the AMIA Symposium*, 105.

P. J. Easterbrook, R. Gopalan, J. A. Berlin, and D. R. Matthews. 1991. Publication bias in clinical research. *Lancet*, 337(8746): 867-872.

M. Egger, G. D. Smith, M. Schneider, and C. Minder. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315(7109): 629-634.

D. Fanelli. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90(3): 891-904.

R. Farkas, V. Vincze, G. Móra, J. Csirik, and G. Szarvas. 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, 1-12.

S. A. Greenberg. 2009. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 339, b2680.

R. S. Hebert, S. M. Wright, R. S. Dittus, and T. A. Elasy. 2002. Prominent medical journals often provide insufficient information to assess the validity of studies with negative results. *Journal of Negative Results in Biomedicine* 1(1):1.

M. Light, X-Y Qiu, and P. 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*, pp. 17-24.

R. Morante, A. Liekens, and W. Daelemans. 2008. Learning the scope of negation in biomedical texts. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 715-724.

R. Morante, and W. Daelemans. 2009. A metalearning approach to processing the scope of negation. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 21-29.

R. Rosenthal. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3): 638.

G. Szarvas, V. Vincze, R. Farkas, and J. 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*, pp. 38-45.