Journal of the American Medical Informatics Association, 25(7), 2018, 855–861
doi: 10.1093/jamia/ocy038
Advance Access Publication Date: 28 April 2018
Research and Applications

Research and Applications

Automatic recognition of self-acknowledged limitations in clinical research literature

Halil Kilicoglu,[1] Graciela Rosemblat,[1] Mario Malički,[2,3] and Gerben ter Riet[2]

[1]Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, USA, [2]Department of General Practice, Academic Medical Center, Amsterdam, The Netherlands and [3]Department of Research in Biomedicine and Health, University of Split School of Medicine, Split, Croatia

Correspondence Author: Halil Kilicoglu, PhD, Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA; kilicogluh@mail.nih.gov

Received 19 September 2017; Revised 21 February 2018; Editorial Decision 25 March 2018; Accepted 28 March 2018

ABSTRACT

Objective: To automatically recognize self-acknowledged limitations in clinical research publications to support efforts in improving research transparency.
Methods: To develop our recognition methods, we used a set of 8431 sentences from 1197 PubMed Central articles. A subset of these sentences was manually annotated for training/testing, and inter-annotator agreement was calculated. We cast the recognition problem as a binary classification task, in which we determine whether a given sentence from a publication discusses self-acknowledged limitations or not. We experimented with three methods: a rule-based approach based on document structure, supervised machine learning, and a semi-supervised method that uses self-training to expand the training set in order to improve classification performance. The machine learning algorithms used were logistic regression (LR) and support vector machines (SVM).
Results: Annotators had good agreement in labeling limitation sentences (Krippendorff's $\alpha = 0.781$). Of the three methods used, the rule-based method yielded the best performance with 91.5% accuracy (95% CI [90.1-92.9]), while self-training with SVM led to a small improvement over fully supervised learning (89.9%, 95% CI [88.4-91.4] vs 89.6%, 95% CI [88.1-91.1]).
Conclusions: The approach presented can be incorporated into the workflows of stakeholders focusing on research transparency to improve reporting of limitations in clinical studies.

Key words: self-acknowledged limitations, clinical research literature, natural language processing, research transparency

INTRODUCTION

In clinical research, study design, execution, and interpretation of results can limit the relevance and applicability of findings. It is important for researchers to acknowledge potential problems and biases as limitations and discuss their magnitude when publishing their findings. This allows stakeholders, such as peer reviewers, journal editors, systematic reviewers, clinicians, and policymakers, to better understand a finding, place it in proper context, assess the potential errors involved, and ascribe to the finding a credibility level.[1] Conversely, failure to discuss limitations (a form of "spin"[2]) may mislead readers to view findings more favorably. Acknowledging limitations can also improve research transparency and reproducibility, thus, reducing research waste.[3]

Despite the importance of limitations in interpreting research findings, authors of biomedical articles often fail to discuss them. A study of 400 articles from six highly-ranked journals showed that only 17% used at least one word denoting limitations.[4] A study on the effect of peer review on manuscript quality found that the acknowledgement of limitations was the most problematic among 34

items of manuscript quality at submission.[5] More recently, ter Riet et al.[6] analyzed 300 biomedical articles and found that 27% did not report limitations. While this is an improvement over earlier reports, it also shows that failure to acknowledge limitations is still a significant problem in the biomedical literature.

Reporting guidelines (e.g., CONSORT,[7] ARRIVE[8]) have been proposed to promote the transparency and accuracy of reporting for biomedical studies, and they often include discussion of limitations as a checklist item. Although such guidelines have been endorsed by high-profile biomedical journals and compliance with them is associated with improved reporting quality,[9] adherence remains sub-optimal.[10]

Text mining tools can automate assessment of a publication by locating key statements corresponding to specific reporting guideline items or alerting of their absence. This can help not only journal editors and peer reviewers in enforcing transparency, but also systematic reviewers who aim to identify rigorous studies and clinicians looking for the best available clinical evidence.[11] Our overarching goal is to develop such text mining tools. In this study, we specifically focus on reporting of limitations, due to its centrality to interpretation of research findings.[4]

Related Work

Studies of self-acknowledged limitations often relied on manual analysis of manuscripts and resulting publications[5,6] or on simple keyword searches.[4]

Information extraction from scientific publications is a well-studied task in biomedical text mining. Among other purposes, information extraction methods have been developed to automate article selection for inclusion in systematic reviews,[12] to identify PICO elements (problem/population, intervention, comparison, and outcome) for evidence-based medicine,[13–16] to recognize key elements of randomized clinical trials,[17] and to extract data deposition statements.[18]

Statistical learning approaches with varying degrees of supervision and rule-based techniques as well as hybrid approaches incorporating both have been explored. For example, rules based on UMLS Metathesaurus[19] concepts and regular expressions have been used to extract population information from abstracts.[13] Fully supervised techniques that rely on annotated data for training have most commonly been explored, and learning algorithms such as SVM, Random Forest, and Conditional Random Fields (CRF) have been proposed.[14–15,17–18] Approaches vary in their task formulation: binary classification,[13,18] multi-label classification,[17] or sequence labeling.[14,18] Commonly used features include *n*-grams, part-of-speech, section, position, and sequence information.

When annotated data is limited, semi-supervised learning approaches have been used to construct somewhat noisy datasets from external resources. Wallace et al.[16] used free-text summaries of PICO elements in the Cochrane Database of Systematic Reviews (CDSR) to automatically generate sentence-level annotations for PICO elements. Marshall et al.[20] used risk of bias judgments in CDSR to semi-automatically generate labels for risk of bias statements in clinical trial publications.

Hybrid approaches that combine statistical learning with rules often work well in practice. For instance, to identify key clinical trial elements (e.g., eligibility criteria, primary outcomes) from relevant publications, Kiritchenko et al.[17] combined an array of sentence-level classifiers with regular expressions to extract the exact phrase for that element.

We are not aware of any biomedical text mining work specifically addressing self-acknowledged limitations. Previous research on rhetorical structure of scientific articles categorized sentences or passages according to their function within this structure.[21–25] The categories identified in such work were often coarse-grained (Aim, Result, etc.), although self-acknowledged limitations were recognized as a *rhetorical move* in the Knowledge Claim Discourse Model.[25] To our knowledge, no automatic recognition approach has been presented for this category.

METHODS

Below, we describe our dataset construction and the methods used to recognize self-acknowledged limitation sentences: a relatively simple rule-based method along with supervised and semi-supervised approaches.

Corpus Construction

We constructed a collection of sentences from PubMed Central full-text articles published in 2017. We followed the approach reported in ter Riet et al.,[6] which ensured a broad selection of papers from general medical and specialty journals. Meta-analyses and systematic reviews were excluded to avoid confusion between the limitations of the included studies and those of the reviews themselves (further selection details are available in the Supplementary Material). The search was conducted on April 12, 2017 and resulted in 1238 articles. XML versions of the articles were downloaded from PubMed Central. Forty-one articles with no usable content (abstract or full-text) were excluded, leaving 1197 articles for further consideration.

For training and testing, we selected a subset of sentences from these articles, based on two observations regarding section structure. First, self-acknowledged limitations often appear in Discussion, Conclusion, or in dedicated sections with the words *limitation* or *weakness* in the title. Second, when not discussed in dedicated sections, limitations can appear in paragraphs introduced by a sentence containing those words (e.g., *Our study has several limitations.*). We used our own rule-based method[26] for sentence splitting and the Stanford CoreNLP toolkit[27] for tokenization.

Based on the observations above, we automatically selected sentences that *potentially* discuss limitations (with some noise) (n_{POS} =2516). We ignored all sentences with fewer than 30 characters, as that often indicated a sentence splitting error. Limitation sentences, if present, constitute a small proportion of all sentences in a full-text article. To generate a somewhat balanced dataset, we limited the number of potentially negative instances selected from an article, as follows:

- If the article was 50 sentences long or shorter, the number of potentially negative instances was set to the number of sentences divided by 10, rounded to the nearest integer less than or equal to the ratio.
- If no potentially positive instances were identified, the maximum number of negative instances was set to 4. Otherwise, the number of negative instances was equal to the number of potentially positive instances.

Given these constraints, a subset of potentially negative instances was randomly selected, resulting in 5915 potentially negative instances (n_{NEG}), for a total of 8431 sentences ($n_{POS}+n_{NEG}$)[1]. We split this dataset into three:

1 This sentence selection process also constitutes our baseline method.

- SEED: 752 sentences (257 potentially positive, 34%) from 98 articles
- TEST: 1505 sentences (479 potentially positive, 32%) from 198 articles
- UNLABELED: 6174 sentences (1780 potentially positive, 29%) from 901 articles

The number of SEED articles reflected the amount of sentence labeling that could realistically be done in a short time (one working day) and was determined based on preliminary annotation performed by one of the authors (HK).

Manual Annotation

Manual annotation was performed on the SEED and TEST sets. Annotation guidelines were developed by two of the authors (HK and GtR) and are provided in the Supplementary Material. Sentences reporting limitations were labeled as POS and those that did not as NEG. Example annotations for three consecutive sentences from an article (PMC5120936) are provided below.

(1) *Our study has a number of limitations.* (NEG)
(2) *First, our findings are limited by the fact that we were unable to enroll our original calculated sample due to slow accrual.* (POS)
(3) *However, an increase in sample size is unlikely to change our findings as an interim review by the DSMB determined that a sample size of 65 using the same assumptions as the original calculation have a power of 70% to 90% for the standard deviation ranging from 0.3 to 0.5.* (NEG)

The first sentence is a negative instance, because it merely announces limitations. The second sentence is a positive instance, since it discusses small sample size as a limitation. The third sentence is also negative, since the authors downplay the significance of the sample size but do not discuss a new limitation.

All four authors annotated the SEED set. One of the authors (GtR) adjudicated the disagreements. Next, three annotators (GR, MM, and GtR) annotated the TEST set to be used for evaluation. Each sentence in this set was double-annotated and then adjudicated by the author who did not participate in annotation of this set (HK). We calculated inter-annotator agreement using Krippendorff's α, which accommodates more than two annotators and missing annotations. We used Microsoft Excel for sentence annotation.

Automatic Recognition of Limitation Sentences

The rule-based method is a refinement of the automatic labeling method used to construct the dataset, and consists of two steps. First, it identifies whether the sentence belongs to a *limitation zone*, and next, it determines whether the sentence is a *limitation-introducing sentence*, *mitigation sentence*, or a *citation sentence*. If the sentence is in a limitation zone, but does not belong to one of these types, it is recognized as a limitation sentence.

A limitation zone is defined as a sequence of sentences that discuss limitations. Our method determines such zones as follows:

- If a section title includes a limitation lemma (*limitation, weakness*) but not a strength lemma (*strength*), the entire section is a limitation zone.
- If the title includes both limitation and strength lemmas, a subsequence in the section is a limitation zone. To determine this subsequence, it first identifies the first sentence containing a limitation lemma (*limitation, weakness*). Then, it identifies the first subsequent sentence containing a strength lemma (*strength*).
  - If both are identified, the limitation zone is the subsequence between these sentences, including the limitation sentence and excluding the strength sentence.
  - If only a limitation sentence is identified, the limitation zone includes this sentence as well as the subsequent sentences to the end of the section.
  - If neither is identified, it checks whether the limitation or strength lemma appears first in the title. If the limitation lemma appears first, it takes the first half of the sentences in the section as the limitation zone. Otherwise, it takes the second half.
- If the title of a section includes either discussion or conclusion, it determines whether any section paragraph begins with a limitation-introducing sentence (described below). If so, the paragraph constitutes a limitation zone.

A limitation-introducing sentence satisfies one of the following constraints:

- It ends with the plural form of a limitation lemma (*limitations, weaknesses*).
- It contains a plural form of a limitation lemma and has 10 tokens or fewer.

To recognize a mitigation sentence, the method first identifies whether the limitation zone itemizes limitations, by searching for a sentence-initial discourse marker that provides this function (*first*, *firstly*, *second*, *secondly*, *third*, *thirdly*, *fourth*, and *fifth*). Next, it ensures that the sentence in question begins with a contrastive marker (*however, nevertheless*, or *nonetheless)*. A citation sentence is recognized with a simple regular expression that detects square brackets and reference numbers.

With these rules, the second sentence in Example (1) is recognized as a limitation sentence. The first and third are non-limitation sentences (limitation-introducing and mitigation, respectively).

Semi-supervised learning with self-training

Semi-supervised learning was used to assess whether we can leverage a small number of manually annotated sentences and a much larger set of unlabeled sentences to obtain good classification performance. If successful, said approach can potentially reduce the burden of time-consuming and labor-intensive annotation and make the development of the envisaged text mining tools more feasible and scalable, since unlabeled data from the biomedical literature can be easily obtained.

Self-training is a semi-supervised learning technique,[28] in which the basic idea is to iteratively choose a proportion of the unlabeled data to augment the labeled data for training to improve accuracy. Given a set of labeled data L and unlabeled data U, a classifier C is trained on L, and U is classified with C. Next, a subset $U' \subset U$ for which C has high confidence is selected. Finally, U' is removed from U and added to L, and the steps are repeated until the algorithm converges. In our case, L is initially the SEED set, and U is the UNLABELED set.

Self-training approaches vary in how the subset U' is determined and when the algorithm is stopped. We established two types of threshold parameters to determine the subset and the stopping criteria. We used an *absolute probability threshold* for each class (α_{POS} and α_{NEG}) and a *probability interval* for the POS class (β_{POS}).

Absolute probability threshold indicates the lowest probability allowed for a given class in any self-training epoch. The probability interval indicates the lowest probability for the POS class in a given epoch. A β_{POS} value of 0.9 indicates that POS probability for an instance has to be within 90% of the probability of the positive instance with the highest confidence to be added to U'. That is, if the highest probability among instances labeled POS is 0.9, a given instance needs to have at least 0.81 probability of being POS to be added to the expanded set. Further, we keep the proportion of POS/NEG instances in the augmented dataset roughly the same as that in the SEED set. Self-training converges when no instances are labeled POS with high enough probability ($Pr(POS) < \alpha_{POS}$), or when all the unlabeled sentences are added to the expanded set.

Self-training can be applied to any supervised learning algorithm. We used support vector machines (SVM) and logistic regression (LR) as learning algorithms, with their LIBLINEAR implementation.[29] The regularization parameter C was set to 1 for both algorithms using grid search. As confidence scores, we used posterior probabilities produced by LR. SVM does not produce probabilities. We obtained probabilities by fitting a standard sigmoid function over scores produced by SVM.

The features we used for learning are lexical and positional features:

- ***n*-grams:** Lemmatized unigrams and bigrams extracted from the sentence.
- **Top section name:** Name of the top-level section in which the sentence appears.
- **Innermost section name:** Name of the innermost subsection to which the sentence belongs.
- **Limitation paragraph:** Binary feature that indicates whether the sentence is in a paragraph that begins with a limitation-introducing sentence.
- **Limitation-strength paragraph:** Binary feature indicating whether the sentence is in a paragraph that begins with a limitation- or strength-introducing sentence. The lemmas *strength* and *strong* were used to identify strength-introducing sentences.

RESULTS

Manual Annotation

Table 1 shows the distribution of manually annotated instances in the SEED and TEST sets.

Krippendorff's α for inter-annotator agreement in the SEED and the TEST sets were 0.774 and 0.789, respectively, bringing the overall agreement to 0.781, considered good agreement.

Limitation Recognition

We evaluated the recognition methods using precision, recall, and F_1 score for the POS class and overall accuracy. We also calculated 95% confidence intervals. We present the main results in Table 2. The baseline method, used to construct the corpus, yielded a reasonable performance (86.4% accuracy). The rule-based method, making several well-targeted enhancements to the baseline, achieved the best accuracy (91.5%). Machine learning in fully supervised mode was less successful (89.5% and 89.6% accuracy with LR and SVM, respectively). Both LR and SVM classifiers yielded comparable precision to the rule-based method, while other metrics were significantly higher for the rule-based method. We obtained higher recall and F_1 score with LR, while SVM yielded the highest precision. Only the recall difference between the two algorithms was statistically significant. Self-training improved SVM performance slightly in all metrics, whereas it led to a drop in precision and accuracy of the LR classifier. We also used the results of the rule-based method to leverage the UNLABELED set. Assigning the labels predicted by the rule-based method to all instances in the UNLABELED set, we obtained the best accuracy overall (91.9% with SVM).

Table 1. Distribution of manually annotated sentences. Each SEED set sentence was annotated by four annotators. Each TEST set sentence was annotated by two out of three annotators

Type	SEED	Pct.	TEST	Pct.	TOTAL	Pct.
Limitation (POS)	164	21.8%	303	20.1%	467	20.7%
Non-limitation (NEG)	588	78.2%	1202	79.9%	1790	79.3%
TOTAL	752		1505		2257	

While the dataset we constructed is more balanced than one based on random sampling of sentences would be, it is still skewed towards negative instances (Table 1). To test whether a more balanced dataset could improve machine learning performance, we undersampled negative examples from the SEED test for training (Table 3). A slightly more balanced set (1: 3 ratio) yielded a minor improvement over the entire SEED set (90.0% vs 89.6%). Moving towards balanced data tended to lower precision while improving recall.

We also experimented with varying the size of the training set. We set aside approximately 20% of the entire annotated dataset (457 instances) as the held out set and varied the number of training examples from 10% of the rest of the dataset (225 instances) to 80% (1800 instances). Increasing the training data size improved the results up to a point (55% in Table 3). Performance with the more traditional 80-20 split was not the highest, though still higher than that obtained with the split reported in Table 2 (shown in italics in Table 3).

DISCUSSION

Our manual annotation confirmed ter Riet et al.'s[6] findings regarding prevalence of self-acknowledged limitation reporting in clinical publications (further dataset characteristics are provided in the Supplementary Material). In clinical publications, self-acknowledged limitations are often highly localized, usually with its own dedicated section or paragraph. Limitation sections/paragraphs are generally a mix of limitation statements and other statements that downplay these limitations or discuss steps taken to mitigate them. Our rules captured this to some extent, achieving good accuracy (91.5%). A common strategy used in scientific writing is to interweave the discussion of strengths and limitations of a study, which can present challenges for automatic approaches.

Machine Learning Performance

The poor performance of machine learning approaches was unexpected. Other statistical learning algorithms we explored (e.g., Random Forest) did not improve performance over those reported here. We also incorporated more sophisticated features used in similar work (e.g., more fine-grained position information, tense, modality) and features based on our rules (e.g., presence of citations, itemized lists, or mitigating sentences), but these had no positive effect either. Various sampling approaches yielded slightly better results. It seems clear that more data does not necessarily lead to better classification, but the question of how much data is sufficient or what data compo-

Table 2. Automatic limitation recognition results (LR = logistic regression, SVM = support vector machines). The 95% confidence intervals are shown in square brackets. All numbers are percentages. For LR, self-training parameters used were α_{POS}=0.7, α_{NEG}=0.95, β_{POS}=0.9. For SVM, they were α_{POS}=0.7, α_{NEG}=0.8, β_{POS}=0.9.

Method	Precision	Recall	F_1 score	Accuracy
Baseline	62.6 [60.2-65.0]	81.2 [79.2-83.2]	70.7 [68.4-73.0]	86.4 [84.7-88.1]
Rules	75.8 [73.6-78.0]	84.8 [83.0-86.6]	80.0 [78.0-82.0]	91.5 [90.1-92.9]
Fully supervised learning with SEED for training				
LR	73.0 [70.8-75.2]	75.9 [73.7-78.1]	74.4 [72.2-76.6]	89.5 [88.0-91.1]
SVM	76.6 [74.5-78.7]	69.3 [67.0-71.6]	72.8 [70.6-75.1]	89.6 [88.1-91.1]
Leveraging UNLABELED for training				
Self-training (LR)	69.4 [67.2-71.8]	84.2 [82.4-86.0]	76.1 [74.0-78.3]	89.4 [87.8-91.0]
Self-training (SVM)	77.1 [75.0-79.2]	71.0 [68.7-73.3]	73.9 [71.7-76.1]	89.9 [88.4-91.4]
Rule-based expansion (LR)	77.4 [75.3-79.5]	81.2 [79.2-83.2]	79.2 [77.2-81.3]	91.4 [90.0-92.8]
Rule-based expansion (SVM)	77.8 [75.7-79.9]	83.5 [81.6-85.4]	80.6 [78.6-82.6]	91.9 [90.5-93.3]

Table 3. Automatic limitation recognition results with varying training data composition and size. The results shown are for the SVM classifier. POS: NEG ratio of 1: 1 indicates a balanced dataset. The SEED-TEST split we used to obtain the results in Table 2 is shown in italics (33.3).

Method	Precision	Recall	F_1 score	Accuracy
Undersampling NEG instances				
POS: NEG Ratio				
1: 1	62.4	89.4	73.5	87.0
1: 2	71.2	79.2	75.0	89.4
1: 3	75.7	73.9	74.8	90.0
Training split size as a proportion of the annotated dataset (SEED+TEST)				
Training Pct.				
10	74.5	67.6	70.9	86.9
20	75.8	69.4	72.5	87.5
30	78.6	71.3	74.8	88.6
33.3	*77.8*	*71.3*	*74.4*	*88.4*
40	78.0	72.2	75.0	88.6
55	81.8	75.0	78.3	90.2
70	83.5	70.4	76.4	89.7
80	83.3	69.4	75.8	89.5

sition is adequate remains, and is likely problem-dependent. To measure the effect of features on classification performance, we conducted an ablation study in which we excluded a distinct set of features from consideration. The results (provided in Supplementary Material) indicated that *n*-gram features were most effective, followed by paragraph features. The effect of section features on performance was smallest.

Overall, the effect of self-training was small, with a slight improvement in SVM accuracy and a slight degradation in LR accuracy. Using smaller initial training sets, self-training with LR also led to a small performance improvement (2.4% accuracy improvement with 150 initial training examples); however, this improvement vanished with larger initial sets (full results provided in Supplementary Material). Whether the improvement obtained with self-training is worth the computational effort remains an open question. With an initial set of 150 examples, self-training with LR converged in 17 epochs and the expanded training set consisted of 5836 examples, for an accuracy improvement of 2.4%. On the other hand, recall improved by 42% at the expense of a small reduction in precision (1.5%). For our purposes, it seems more important not to miss limitation sentences; thus, a small loss in precision with a significant improvement in recall may justify the additional self-training effort. In self-training, the models resulting from intermediary epochs sometimes yielded better performance on the test set; therefore, it might be possible to design better stopping criteria. Using the labels predicted by the rule-based method for expansion yielded best overall performance, indicating that the rules we devised were robust. Furthermore, this rule-based expansion represents an alternative approach to labeling data at a low cost (with some noise), assuming a simple rule-based approach can be devised for the task under consideration. In addition to the threshold-based approach presented here, we also experimented with another, more sophisticated self-training approach (*self-training with selection-by-rejection*[30]) which did not improve performance.

Error Analysis

Analyzing the errors made by the rule-based method, we found that some of our assumptions were not reflected in some articles, leading to false positive errors. For example, we assumed that in a "strength and limitation" section, every sentence addressed a strength or a limitation. In the following passage, which concludes such a section (PMC5120917), no specific strength or limitation is discussed.

(2) *In conclusion, the present study shows that women with UL are at greater risk of prevalence of MetS regardless of confounding factors. We suggest the biological mechanism responsible for UL may involve IR aggravation, . . .*

In some cases, we were unable to determine that the limitations discussed were those of other studies. An example that led to several false positive errors is given below (PMC5120938):

(3) *Although several studies have demonstrated the beneficial effect of portal hyperperfusion on hepatic regeneration in LDLT recipients, these studies have certain limitations. . .*

Such statements can be distinguished by the presence of phrases like *these studies*. Owing to the incremental nature of rule-based methods, it would be feasible to add such a rule to improve performance.

We missed some limitation sentences due to the presence of citations, which we interpreted as indicating a limitation of another study. An example is given below (PMC5120923):

(4) *First, the scale that was used in our study might be a crude measure, although it has been validated for evaluating occupational stress [10].*

A small number of limitations appeared outside a specific limitation zone, while in other cases, our inability to identify the limitation-introducing statement led to false negative errors. In the example below (PMC5120933), the second sentence was not recognized as a limitation.

(5) *This study has intrinsic limitations and pitfalls in the diagnosis, outcome, and patient population. . . . . For these reasons, the incidence of AIHA may have been underestimated.*

We also found that the triggers we used for itemized lists (*first*, *finally*, etc.) can be expanded to include words like *additionally* and *moreover* for improved accuracy.

Limitations of Our Study

Our study has several limitations. While we aimed for a fair sample of articles, our dataset may not be representative of the clinical research literature. Our sentence selection method was the same as the baseline method, which might have introduced bias towards "positive" sentences, although we tried to minimize this bias by manual sentence annotation. We hand-crafted features for machine learning based on our observations and intuitions; however, these features may not be optimal. Exploring representation learning[31] (i.e., automatic discovery of representations needed for feature detection) could prove fruitful. Self-training is one of the simplest semi-supervised learning methods. More sophisticated approaches[28] could yield more significant performance improvements.

CONCLUSION

We presented an annotated dataset of self-acknowledged limitation sentences and developed automatic methods to recognize such statements. Our rule-based method, despite its relative simplicity, outperformed the supervised machine learning methods. Self-training yielded minor differences over fully supervised methods. While the performance of the rule-based method may be somewhat study-specific, it also suggests that such methods, often dismissed as being difficult to develop and brittle, should not be assumed inferior to the statistical learning methods. Depending on the problem, they can lead to more parsimonious models with greater predictive power, using a fraction of the training data. The dataset resulting from this study is publicly available (https://doi.org/10.5061/dryad.06ds7) and the rule-based method is available upon request.

There are several future directions for this study. Recognizing limitation sentences is a useful first step toward recognition and classification of individual limitations stated in these sentences, which could be used to generate a limitation profile of a publication. Such work depends on the construction of a taxonomy of limitation types (e.g., limitations of internal validity, such as measurement errors, and those of external validity, such as selected study populations[6]) We also plan to apply our methodology at a large scale to biomedical articles, which would benefit applications that focus on reporting rigor and transparency.

FUNDING

This work was in part supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

CONTRIBUTORS

HK and GtR conceived of the study and developed the annotation guidelines. HK, GR, MM, and GtR contributed to annotation. HK led the annotation, managed the data, performed calculations and analyses, and drafted the manuscript. All authors read and approved the manuscript.

COMPETING INTERESTS

None.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2005; 2 (8): e124.
2. Chiu K, Grundy Q, Bero L. Spin'in published biomedical literature: a methodological systematic review. *PLoS Biol* 2017; 15 (9): e2002173.
3. Munafò MR, Nosek BA, Bishop DVM, *et al*. A manifesto for reproducible science. *Nat Hum Behav* 2017; 1 (1): 21.
4. Ioannidis JPA. Limitations are not properly acknowledged in the scientific literature. *J Clin Epidemiol* 2007; 60 (4): 324–9.
5. Goodman SN, Berlin J, Fletcher SW, Fletcher RH. Manuscript quality before and after peer review and editing at Annals of Internal Medicine. *Ann Intern Med* 1994; 121 (1): 11–21.
6. ter Riet G, Chesley P, Gross AG, *et al*. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. *PLoS One* 2013; 8 (11): e73623–6.
7. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340: c332.
8. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010; 8(6): e1000412.
9. Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. *J Clin Epidemiol* 2007; 60 (3): 241–9.
10. Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst Rev* 2012; 1 (1): 60.
11. Kilicoglu H. Biomedical text mining for research rigor and integrity: tasks, challenges, directions. *Brief Bioinform* 2017; bbx057. doi:10.1093/bib/bbx057.
12. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev* 2015; 4 (1): 5.
13. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. *Comput Linguist* 2007; 33 (1): 63–103.
14. Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics* 2011; 12 (Suppl 2): S5.

15. Hassanzadeh H, Groza T, Hunter J. Identifying scientific artefacts in biomedical literature: the Evidence Based Medicine use case. *J Biomed Inform* 2014; 49: 159–70.
16. Wallace BC, Kuiper J, Sharma A, Zhu M, Marshall IJ. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *J Mach Learn Res* 2016; 17 (132): 1–25.
17. Kiritchenko S, de Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak* 2010; 10 (1): 56.
18. Névéol A, Wilbur WJ, Lu Z. Extraction of data deposition statements from the literature. *Bioinformatics* 2011; 27 (23): 3306–12.
19. Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. *Methods Inf Med* 1993; 32 (04): 281–91.
20. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc* 2016; 23 (1): 193–201.
21. Teufel S, Carletta J, Moens M. An annotation scheme for discourse-level argumentation in research articles. In: proceedings of EACL. Stroudsburg, PA: Association of Computational Linguistics (ACL); 1999. p. 110–117.
22. Teufel S, Siddharthan A, Batchelor CR. Towards domain-independent argumentative zoning: evidence from chemistry and computational linguistics. In: proceedings of EMNLP. Singapore: Association of Computational Linguistics (ACL); 2009. p. 1493–1502.
23. Agarwal S, Yu H. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics* 2009; 25 (23): 3174–80.
24. Liakata M, Teufel S, Siddhartan A, Batchelor C. Corpora for conceptualisation and zoning of scientific papers. In: proceedings of LREC 2010. Valletta, Malta: European Language Resources Association (ELRA); 2010. p. 2054–2061.
25. Teufel S. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Stanford, CA: Center for the Study of Language and Information (CSLI); 2010.
26. Kilicoglu H, Demner-Fushman D. Bio-SCoRes: a smorgasbord architecture for coreference resolution in biomedical text. *PLoS One* 2016; 11 (3): e0148538.
27. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, MD: Association of Computational Linguistics (ACL); 2014. p. 55–60.
28. Zhu X. *Semi-Supervised Learning Literature Survey*. Computer Sciences, University of Wisconsin-Madison; 2005: 1530.
29. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 2008; 9: 1871–4.
30. Zhou Y, Kantarcioglu M, Thuraisingham B. Self-training with selection-by-rejection. In: proceedings of IEEE 12th International Conference on Data Mining (ICDM). Washington, DC: IEEE; 2012. p. 795–803.
31. Bengio Y, Courville A, Vincent P. *Representation Learning: A Review and New Perspectives*. Université de Montréal; 2012. Available from: http://arxiv.org/abs/1206.5538.